Universität Basel

Fakultät für Psychologie

Assessments in School Psychology:

Comparability and Diagnostic Utility of Tests Measuring Cognitive Abilities

Inauguraldissertation zur Erlangung der Würde einer Doktorin der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Anette Bünger

aus Lüsslingen SO, Schweiz

Basel, 2020

Universität Basel

Fakultät für Psychologie

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Alexander Grob

Prof. Dr. Claudia M. Roebers

Datum des Doktoratsexamen: 24. März 2020

DekanIn der Fakultät für Psychologie

Universität Basel

Fakultät für Psychologie

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

- **Bünger, A.**, Grieder, S., Schweizer, F., & Grob, A. (2019). The comparability of intelligence test results: Group- and individual-level comparisons of seven intelligence tests. *Manuscript submitted to the Journal of School Psychology*.
- Grieder, S., **Bünger, A.**, Odermatt, S. D., Schweizer, F., & Grob, A. (2019). Limited individual-level comparability of IQs within three test batteries: Impact on external validity, some explanations, and possible solutions. *Manuscript submitted to the Journal of Educational Psychology.*
- **Bünger, A.**, Urfer-Maurer, N., & Grob, A. (2019). Multimethod assessment of attention, executive functions, and motor skills in children with and without ADHD: Children's performance and parents' perceptions. *Journal of Attention Disorders.* Advance online publication. doi.org/10.1177/108705471882498

Basel, den 22.01.2020

Anette Bünger

ACKNOWLEDGMENTS

With lots of gratitude to the following people for their support throughout my PhD:

To Prof. Dr. Alexander Grob for the guidance, the continuous support, and the confidence you have shown in me throughout my PhD.

To Prof. Dr. Claudia Roebers for the continuous support and encouragement, and for the many opportunities you have provided me from the moment we met.

To Prof. Dr. Rainer Greifeneder for serving on the dissertation committee.

To the Suzanne and Hans Biaesch Foundation for the Enhancement of Applied Psychology for placing their trust in me and for having awarded me a grant to conduct my dissertation project.

To my current and former PEP- and SEED-colleagues for the fabulous team spirit and the enriching exchange of knowledge. A special thanks to Dr. Wenke Möhring, Silvia Grieder, and Salome Odermatt for your thoughtful comments on previous versions of this dissertation.

To my appreciated colleagues from the School Psychology Service Basel-Stadt for your open doors, your encouragement, and for your genuine interest in my work.

To my partner Sascha Michel for your support in countless ways, for making each day a bit brighter, and for inspiring me with your light-footed way of living.

To my parents Annemarie and Martin Bünger-Suter for your unconditional love and support throughout my whole life. With open minds you taught me how to look at things from different perspectives. To you I dedicate this thesis.

TABLE OF CONTENTS

ABSTRACT

In the context of school psychology, students' cognitive abilities such as intelligence, attention, and executive functions are often assessed for diagnostic classification purposes (e.g., intellectual disability or giftedness; the presence of an attention-deficit/hyperactivity disorder [ADHD]). There are an increasing number of tests measuring these abilities, raising the question of whether different tests aimed at measuring the same psychological attribute yield comparable results. Most previous studies have addressed this question using group-level analyses (i.e., correlations among tests). Yet, research investigating whether different tests yield comparable scores for individuals, that is, scores that lead to the same classification, is sparse. This is surprising given that individual-level comparability is essential, as diagnostic classification is most often based on cut-off scores. Thus, the main goal of the present dissertation was to shed light on such individual-level comparability with findings from three studies. Studies I and II revealed that different intelligence test scores were generally highly correlated on the group level, but individual-level comparability was not satisfactory. Specifically, in Study I, the 95% confidence interval (CI) of two test scores obtained from *different* tests overlapped in only about 60% of all cases. In Study II, the 95% CI of intelligence scores obtained from the *same* test (Full-Scale IQs vs. Screening IQs) did overlap in 74–99% of all cases. In both studies, comparability decreased toward the tails of the IQ distribution, the very ranges in which diagnostic questions most often arise in practice. Study III revealed that scores of attention and executive functions obtained in parent questionnaires did not correlate substantially with those obtained in performance-based measures. Moreover, only scores from parent questionnaires and not those from performance-based measures were associated with an ADHD diagnosis. Possible approaches for dealing with and enhancing individual-level comparability are discussed. The focus thereby lies on two aspects that were identified to be the most prominent sources of incomparability: different theoretical groundings of tests and measurement error.

Introduction

In the field of school psychology, the abilities of children and adolescents are frequently assessed following two general goals: One is the identification of strengths and difficulties of students and *how* best to support them during their school career. The second purpose is the *classification* of students' abilities into discrete categories (e.g., intellectually disabled, normal, or gifted; criteria for a diagnosis are met or not met) to determine if a child is eligible for special education. The second purpose is often necessary because the nature of diagnoses is discrete or dichotomous (American Psychiatric Association, 2013; World Health Organization [WHO], 2018) and because public policies often bind financial resource allocation for special education to the presence of a diagnosis. Yet, it may raise an ethical conflict in psychologists, as they know that behavior is complex and falls on a continuum (Weissman & Debow, 2003). This dilemma can be eased by using quantitative, age-standardized psychometric test procedures such as intelligence tests, neuropsychological test batteries, and standardized behavior-rating scales. Such tests provide unambiguous scores that facilitate the categorization of a child or adolescent's abilities according to prespecified (quantitative) criteria and cut-off scores, which in turn increase the objectivity of diagnostic decisions that follow a test score. Whereas there is no universal definition of abnormal test performance, it is often referred to as the standard deviation from the mean in order to define a strength or an impairment (e.g., Binder, Iverson, & Brooks, 2009; Heyanka, Holster, & Golden, 2013). By using a standardized test and cut-off scores, the psychologist's responsibility is slightly shifted away from the classification itself toward the process of selecting an appropriate test, administering it correctly, and interpreting results according to the current state of research and test standards (Diagnostik- und Testkuratorium, 2018; International Test Commission, 2001) and in line with ethical principles (e.g., American Psychological Association, 2017). Although not binding, these guidelines inter alia ask test developers to provide evidence of the test's psychometric properties (e.g., objectivity,

reliability, and validity), and they ask test administrators to select, apply, and interpret tests correctly and with consideration of their limits (e.g., restricted reliability or validity). To act accordingly, school psychologists must be familiar or, even better, experienced with the tests at their disposal. Yet, the number and diversity of available tests has drastically increased, making it challenging for school psychologists to be familiar with or have access to them all. An essential question therefore is whether different tests that aim at measuring the same psychological attribute yield comparable results for individuals. Especially when psychometric test scores are used for classification, such comparability is vital to ensure that the classification or a given diagnosis depends on an individual's true ability rather than on test selection or any other construct-irrelevant variance (i.e., measurement error).

There is a large body of research investigating how tests measuring the same psychological attribute correlate, but such group-level analyses cannot be directly transferred to the individual level. That is, a high correlation of two test scores does not imply that these test scores will result in the same classification and diagnostic conclusion. Hence, research is needed to gain a more profound, multifaceted, and practically relevant understanding of the comparability of test scores. The present research therefore was aimed at investigating whether psychometric tests that purport to measure the same psychological attribute yield comparable results for individuals, that is, results that lead to the same classification. Although this is relevant to the measurement of all psychological attributes, the focus of the three studies included in this dissertation lies on the comparability of tests measuring cognitive abilities, that is, intelligence (Studies I and II), as well as attention and executive functions (Study III). These particular cognitive abilities were chosen because of their high relevance for real-life outcomes (e.g., academic and occupational success) as well as their pivotal role in school psychology (Benson et al., 2019; Moffitt et al., 2011; Roth et al., 2015). To embed the three studies theoretically, the role of intelligence, attention, and executive functions in school psychology as well as the most prominent construct theories are

highlighted first. Second, characteristics that affect measurement accuracy and that may vary across tests measuring the same attribute, hence threatening test comparability, are revealed. Results from the three studies are presented and discussed. They form the basis of a general discussion on how practitioners and researchers can deal with limited individual-level comparability and how this comparability might be increased in the long term.

Theoretical Background

General intelligence is defined as a broad mental capacity to reason, solve problems, comprehend complex ideas, and learn quickly (Gottfredson, 1997). Binet and Simon (1904) developed the first major and individually administered intelligence test over a century ago to predict interindividual differences in scholastic achievement. Since then, the frequency and diagnostic purposes of intelligence assessment as well as the number and diversity of available measurement instruments have increased drastically (S. Goldstein, Princiotta, & Naglieri, 2015). Most well-established tests are individually administered, and performance based and provide age-standardized general intelligence scores (i.e., IQs). Today, intelligence is the most frequently assessed attribute among school psychologists (Benson et al., 2019), This may be explained by the fact, that numerous studies have shown the association between IQ and educational success (Deary, Strand, Smith, & Fernandes, 2007; Roth et al., 2015). Furthermore, many diagnoses that are relevant in the field of school psychology can only be given in relation to a standardized intelligence test score. In this context, intelligence is assessed to identify intellectual disability or giftedness, which entitles an individual to special education (Bergeron, Floyd, & Shands, 2008; Swiss Conference of Cantonal Ministers of Education, 2007). Intellectual disability or giftedness is defined as two or more standard deviations below or above the mean, respectively, obtained on normed and standardized tests. The most current versions of the *International Classification of Diseases for Mortality and Morbidity Statistic* (*ICD-11*; WHO, 2018) and the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association, 2013) further require ruling out

the presence of an intellectual disability as an alternative explanation for symptoms associated with certain developmental disorders such as attention-deficit/hyperactivity disorder (ADHD), language, or learning disorders. Yet, the *ICD-11* will not come into effect until January 2022. The current valid version, the *ICD-10*, even requires a significant discrepancy (usually 1.2 to 1.5 *SD*s) between below-average school performance (e.g., reading, writing, or mathematical skills) and IQ for the diagnosis of a learning disorder (WHO, 1990). It is therefore crucial to have very accurate intelligence measures to ensure that the presence or absence of a diagnosis does not vary with the use of different tests.

Many intelligence researchers have strived to meet this challenge: The field of (psychometric) intelligence is one of the most prominent and successful research fields in psychology (Stern & Neubauer, 2016). Nevertheless, controversial debates and theories about the psychometric structure of intelligence have persisted (e.g., Beaujean & Benson, 2019a, 2019b; Gignac, 2008; Kovacs & Conway, 2019; Stern & Neubauer, 2016). Major controversies are whether the general intelligence factor truly exists, whether intelligence is unidimensional (e.g., two-factor theory; Spearman, 1927) or multidimensional (e.g., gf-gc theory; Cattell, 1987), and if it is multidimensional, what type of hierarchical model best represents the structure of abilities. Nevertheless, most contemporary tests give a nod to the Cattel–Horn–Carroll (CHC) theory of cognitive abilities (McGrew, 1997, 2005, 2009) to provide a theoretical grounding. CHC theory integrates theories from Cattell (1941), Horn and Cattell (1966), and Carroll (1993). In CHC theory, intelligence is modeled as a multidimensional construct consisting of a number of different abilities that are hierarchically structured on three strata: A general intelligence factor *g* is on the top Stratum III, broad abilities are on Stratum II, and more narrow abilities are on Stratum I. Tests relating to CHC theory usually provide not only a Full-Scale IQ (FSIQ) representing *g* but also a varying number of factor index scores (group factors) representing the more specific broad abilities and subtests representing more narrow abilities. The fact that most contemporary intelligence

tests relate to CHC theory strengthens the assumption of interchangeability of IQs that are derived from different test batteries. However, none of the tests include the measurement of all broad and narrow abilities. Therefore, the specific operationalizations of general intelligence still differ among tests (Beaujean & Benson, 2019b). That is, IQs are calculated based on different sets of subtests that vary in number, task format, heterogeneity of content, and their *g* loadings. These characteristics of composite scores have an influence on the accuracy of composites (Farmer, Floyd, Reynolds, & Berlin, 2019; Jensen & Weng, 1994). According to a thorough investigation by Farmer et al., the most accurate composites are those derived from numerous (at least 4 but up to 12) highly *g*-loaded and diverse subtests. Especially for IQs that are based on fewer than four subtests, inadequate content sampling poses a risk of loss of accuracy of IQs and may therefore reduce comparability (Farmer et al., 2019; Floyd, Clark, & Shadish, 2008). Nevertheless, as time efficiency plays an important role in diagnostic contexts, an increasing number of intelligence tests provide a Screening IQ (ScrIQ) as an index for general intelligence in addition to the FSIQ. The ScrIQ is usually composed of only two subtests that target the broad abilities of fluid and crystallized intelligence (Grob, Gygi, & Hagmann-von Arx, 2019; Grob & Hagmann-von Arx, 2018; Hagmann-von Arx & Grob, 2014).

Numerous studies have confirmed that IQs derived from different intelligence test batteries as well as FSIQs and ScrIQs from the same test battery are highly correlated (e.g., Allen, Stolberg, Thaler, Sutton, & Mayfield, 2014; Baum, Shear, Howe, & Bishop, 2015; Grob & Hagmann-von Arx, 2018; Hagmann-von Arx & Grob, 2014). This is seen as an indication of convergent validity, that is, that the test scores represent the same constructs (Neukrug & Fawcett, 2014). However, as described above, this does not automatically lead to the conclusion that two different tests measuring general intelligence would render comparable results for the same individual. Only two studies so far have addressed this question of individual-level comparability by investigating in how many cases the 90%

confidence intervals (CIs) of FSIQs derived from different tests overlapped. These studies found overlap in 64–76% when investigating schoolchildren (Floyd et al., 2008; Hagmann-von Arx, Lemola, & Grob, 2018), implying that decisions and diagnoses based on a single test are of questionable validity for a fourth to a third of tested children. As this is highly relevant for the validity of test diagnostic, it is crucial to better understand what threatens individual-level comparability of intelligence scores and how it might be enhanced. To this end, investigations are needed of individual-level comparability for the whole age range in which school psychology questions arise (age 4 to 20 years), not only for IQs and factor scores obtained in different tests, but also for FSIQs and ScrIQs obtained within the same test.

Unlike intelligence, the constructs of attention and executive functions have only recently gained increased consideration by school psychologists. Different theories exist to describe the construct of attention (Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991; Posner & Rothbart, 2007; van Zomeren & Brouwer, 1994). These theories differ with regard to the specific attention components that are included in their models (e.g., alertness, sustained and executive attention). Yet, they have in common that attention is modeled as a multicomponent construct that contains bottom-up as well as top-down processes (Katsuki & Constantinidis, 2014; Zelazo et al., 2013). The latter show great overlap with the construct of executive functions, which are widely understood to be a heterogeneous set of higher order cognitive processes that are crucial for goal-directed, purposeful behavior or nonhabitual responses in novel, complex, and challenging situations. Classic components of executive functions are updating, inhibition, and shifting of attention (Miyake et al., 2000).

Whereas these theories mainly focus on cognitive aspects, recent theories include attentional processes as well as executive functions as part of a more comprehensive system of self-regulation that encompasses a complex and entangled set of biological, emotional, behavioral, and cognitive control and regulation processes (Blair & Raver, 2012, 2015; Calkins & Williford, 2009). The increased consideration of attention and executive functions

in school psychology might be explained by three main circumstances: First, these constructs were shown to be substantially associated with developmental aspects that encompass not only school readiness, learning, and academic achievement but also physical and mental health later in adulthood (Best, Miller, & Naglieri, 2011; Blair & Razza, 2007; Moffitt et al., 2011; Roebers, Röthlisberger, Cimeli, Michel, & Neuenschwander, 2011). Second, intervention studies have found promising effects, suggesting that attention and executive functions are malleable and may be improved with specific trainings (Diamond, 2012). Third, deficits of attention and executive functions are core symptoms of ADHD, a neurodevelopmental disorder with a relatively high and stable prevalence of 5% (Polanczyk, Willcutt, Salum, Kieling, & Rohde, 2014). This disorder is persistent across settings and is associated with various functional impairments, including learning problems, scholastic underachievement, and social withdrawal (Barkley, 1997; Maedgen & Carlson, 2000; Polanczyk et al., 2014; Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005). Meeting criteria for an ADHD diagnosis makes children eligible for special education in Switzerland and many other countries worldwide (Polanczyk et al., 2014; Swiss Conference of Cantonal Ministers of Education, 2007). Hence, it is crucial to develop tests that measure attention and executive functions accurately, that is, reliably and validly. Yet, because the processes of attention, executive functions, and self-regulation are intertwined, it is difficult to disentangle them for measurement and diagnostic purposes. Therefore, performance-based tests often include tasks or questions measuring both attention and executive functions (Drechsler, 2007). Furthermore, questionnaires or rating scales that measure these constructs often include emotional and behavioral aspects of self-regulation as well (e.g., Gioia, Isquith, Guy, & Kenworthy, 2000). Hence, other than in the realm of intelligence, different tests measuring attention and executive functions are seen as heterogeneous. They vary with regard to measurement type or method (e.g., performance-based or questionnaires), content (e.g., cognitive, behavioral, emotional), and type of index score they provide (a general attention or

executive function index, analogous to the FSIQ, and/or specific ability index scores, analogous to factor indices or single subtest scores in intelligence tests). As a consequence of these variations, correlations among different tests are moderate for the same measurement types (Caye, Machado, & Rohde, 2017; van der Ven, Kroesbergen, Boom, & Leseman, 2013) and rather low for different measurement types (Toplak, West, & Stanovich, 2013). However, there are studies showing that questionnaires as well as performance-based measures are useful for identifying impairments in children with ADHD (Fried, Hirshfeld-Becker, Petty, Batchelder, & Biederman, 2015; Gioia, Isquith, Kenworthy, & Barton, 2002) Yet, in the context of classifying a child or adolescent with ADHD, it is crucial to know if different types of tests reveal a suspected impairment to a similar degree in the same sample.

Like all psychological attributes, intelligence, attention, and executive functions are latent constructs that are not directly observable. This has two implications that are relevant for test comparability: First, test developers must rely on a specific theoretical model for test construction. As shown above, these theoretical models, or the operationalization of the same theoretical model, may vary across tests and therefore limit comparability of test scores. Second, psychometric test scores, whether they stem from performance-based tests or rating scales, are approximations and not pure representations of an individual's true (theoretically conceptualized) ability. They therefore always contain measurement error (e.g., Shultz, Whitney, & Zickar, 2014).

Measurement error limits a test's accuracy (i.e., reliability and validity) and thus the comparability across test scores. Sources of measurement error are manifold. They may stem from limitations in test construction or test administration. It is assumed that measurement error can be reduced, yet not eliminated, by following test standards and guidelines (National Research Council, 2002). Therefore, test manuals should always contain reliability coefficients as indicator of measurement accuracy (Diagnostik- und Testkuratorium, 2018; International Test Commission, 2001). Reliability coefficients are directly relevant for test use

and interpretation, as they build the basis for calculating a test's standard error of measurement, which is essential for the optional calculation of the CI around a standardized test score. The lower the reliability coefficient is, the wider the CI gets. According to Evers (2001a), psychologists should use tests that have reliability coefficients of at least .90 when making high-stakes diagnostic decisions. However, there are different types of reliability coefficients, each estimating a different source of measurement error. The most common reliability coefficients are (a) internal consistencies (e.g., Cronbach's α or split-half reliability), which consider error of content sampling that is associated with a weak theoretical grounding (Beaujean & Benson, 2019b; Farmer et al., 2019), (b) test–retest reliability coefficients, which consider transient error, that is, temporary states of an individual (e.g., changes in mood or information processing efficiency over time), as well as environmental random circumstances (Schmidt, Le, & Ilies, 2003), and (c) interrater reliability coefficients, which consider examiner influences that may interfere with an examinee's test performance or rating result (e.g., inappropriate help, different expectations of or biased perspectives on an individual's performance; Caye et al., 2017; McDermott, Watkins, & Rhoad, 2014).

Yet there are also sources of measurement error that cannot be captured by these reliability coefficients. Such measurement error might, for instance, stem from different theoretical groundings that direct content sampling (Beaujean & Benson, 2019b; Shultz et al., 2014). In addition, tests may differ in standardization samples and therefore produce culture or cohort effects. Moreover, differences in task format (e.g., verbal or nonverbal, paper-and-pencil or computer based; Floyd et al., 2008) may interact with conditions and characteristics of individual examinees (e.g., language, or motor skills) and hence cause construct-irrelevant variability that differs between tests. If these sources of measurement error are not taken into account when building or interpreting test scores, they may lead to misdiagnosis.

The majority of well-established intelligence tests provide a 95% CI for FSIQs and ScrIQs as well as factor scores. The CI is usually based on overall internal consistency

coefficients, which generally exceed the desired minimum of .90. By contrast, reliability coefficients of attention and executive function tests are often lower than .90 (e.g., Soveri, Lehtonen, Karlsson, Lukasik, & Antfolk, 2018; Syväoja et al., 2015). Nevertheless, many tests measuring attention and executive functions do not provide CIs. When provided, the CI should be used for test interpretation instead of the exact test score, as it takes measurement error into account. Yet, it should be kept in mind that even with the interpretation of CIs, only one of many types of measurement error is considered (e.g., content sampling).

To sum up, different theoretical assumption as well as measurement error may limit the comparability of test scores, especially if the latter is not adequately taken into account by using reliability coefficients and interpreting CIs instead of the exact test score. Because group-level comparisons do not directly transfer to individual, diagnostic settings, more research is needed to better understand to what extent test scores representing the same psychological attribute render comparable results for individuals.

The Studies

The overarching goal of this research was to gain a more profound understanding of comparability among test scores purporting to represent the same psychological attribute. The focus was on intelligence, attention, and executive functions given their relevance for school psychology. The present dissertation encompasses three studies that shed light on this question from different perspectives and with different study designs. Studies I and II aimed at investigating individual-level comparability of different intelligence scores in mainly typically developing individuals. For this purpose, we used two different approaches: the first was to examine whether different IQs of an individual lie within the same classification range (i.e., nominal intelligence level such as average, below or above average). The second was to examine whether CIs plotted around the different IQs overlapped. The CIs were calculated on the basis of the overall internal consistency coefficients (i.e., Cronbach's α or split-half reliability, as indicated in the test manuals) with the standard error of estimate and the

estimated true score, which entails a correction toward the mean. Furthermore, possible predictors of IQ differences (e.g., age, native language, IQ level, and order of test administration) were investigated in both studies. In Study II, we additionally varied the reliability coefficients used for the calculation of CIs to explore whether comparability can be enhanced by taking other sources of measurement error into account. The main distinction between the two studies was that in Study I, we investigated comparability of FSIQs and verbal and nonverbal factor scores among *different* test batteries (between-test comparison), and in Study II we contrasted FSIQs and ScrIQs from the *same* test battery (within-test comparison).

Study III focused on attention and executive functions and differed from the other studies not only with regard to the investigated constructs but also with regard to the study design: Given that attention and executive functions are assessed via either questionnaires (rating scales) or performance-based measures, we investigated the comparability of the two measurement methods. Whereas intelligence tests usually differentiate quite well across the entire IQ distribution and can be applied for detecting difficulties as well as strengths, tests of attention and executive functions are often constructed to detect difficulties and likely show a ceiling effect with typically developing samples. Therefore, we investigated children with an expected deficit in attention and executive functions, namely, children with ADHD, and a control group of typically developing children who were matched with regard to age and gender. The main question was to investigate whether both measurement methods would reveal the expected difficulties in the ADHD group. Because the investigated measures of attention and executive functions do not provide a CI for the obtained score, we did not consider the overlap of CI as an index of comparability. Last, impairments of motor skills were further investigated in Study III, because of their high comorbidity in ADHD. Yet, motor skills are not the focus of the present thesis and will not be further discussed in the following.

Summary of Study Results

In Study I "The comparability of intelligence test results: Group- and individual-level comparisons of seven intelligence tests", we investigated whether participants aged 4–20 years obtained comparable IQs on different well-established intelligence tests, that is, the Intelligence and Development Scales–2 (IDS-2; Grob & Hagmann-von Arx, 2018) and the German adaptations of the Stanford–Binet Intelligence Scales–Fifth Edition (SB5; Grob et al., 2019), the Reynolds Intellectual Assessment Scales (RIAS; Hagmann-von Arx & Grob, 2014), the Snijders-Oomen Nonverbal Intelligence Test 6-40 (Tellegen, Laros, & Petermann, 2012), the Wechsler Adult Intelligence Scale–Third Edition (von Aster, Neubauer, & Horn, 2006), the Wechsler Intelligence Scales for Children–Fourth Edition (Petermann & Petermann, 2011), and the Wechsler Preschool and Primary Scale of Intelligence–Third Edition (Petermann, 2009). All of these tests are either explicitly based on or relate to CHC theory. As expected, findings indicate mostly substantial correlations between IQs. When investigating whether the specific IQs of each individual fell within the same nominal category (e.g., below or above average) or whether the CIs of their obtained IQs overlapped, we found that this was true overall for only 62% of all comparisons. IQs tended to be higher for tests with older standardization years. We found no systematic effect of age or native language on IQ differences. Yet, comparability decreased toward the tails of the IQ distribution. That is, for individuals who scored in the above- or below-average range on at least one intelligence test, comparability decreased to less than 50% on both criteria (overlap of CIs and correspondence of nominal category). Similar patterns appeared for the individual-level comparability of verbal and nonverbal intelligence factor scores.

In Study II "Limited individual-level comparability of IQs within three test batteries: Impact on external validity, some explanations, and possible solutions", we investigated whether FSIQs and ScrIQs obtained from the same test battery were comparable on the group and individual level. The included test batteries were the IDS-2, RIAS, and SB5, which

provide FSIQs as well as ScrIQs and in total cover a large age span from early childhood to late adulthood. We found that despite high group-level comparability (correlations and mean differences), individual-level comparability (correspondence of nominal category and overlapping CIs) was not always satisfactory, especially for younger children as well as toward the tails of the IQ distribution. When differences between the FSIQ and ScrIQ were high, FSIQs tended to predict school grades better than ScrIQs. Last, when internal CIs were calculated on the basis of either age- and IQ-specific internal consistency or test–retest reliability coefficients instead of overall internal consistency coefficients, CIs were enlarged from 6–15 IQ points to 8–30 IQ points. Accordingly, the overlap of CIs increased from an average of 74–99% to an average of 94–100%.

In Study III "Multimethod assessment of attention, executive functions, and motor skills in children with and without ADHD: Children's performance and parents' perceptions", our main focus was to investigate whether parents' perceptions assessed with questionnaires (Conners 3; Lidzba, Christiansen, & Drechsler, 2013) and performance-based measures (CANTAB; Fray, Robbins, & Sahakian, 1996) revealed comparable results when assessing attention and executive functions in children with and without ADHD aged 6–13 years. Results revealed that parent questionnaires and performance-based measures were barely comparable. That is, the two measurement methods did not correlate significantly. Moreover, parent questionnaires but not performance-based measures revealed lower mean values and a higher number of children showing an impairment in the ADHD group. Parent-reported difficulties but not performance-based scores of attention and executive functions were related to an ADHD diagnosis.

Discussion

The goal of the three studies included in this dissertation was to gain a more comprehensive knowledge of the comparability of test scores representing intelligence, attention, and executive functions. Studies I and II had a similar study design and were in fact

consecutive. Therefore, findings from these two studies are discussed first, before integrating results of all three studies to derive options for research and practice to deal with or enhance comparability of test scores.

Studies I and II both revealed that despite high correlations of intelligence scores, comparability on the individual level (correspondence of nominal category and overlap of CIs) was not satisfactory. Moreover, the magnitude of IQ differences and individual-level comparability was lower toward the tails of the IQ distribution. This was found when comparing FSIQs and nonverbal and verbal factor indexes from different test batteries (Study I) as well as when comparing FSIQs and ScrIQs from the same test battery (Study II). Notably, the tails of the IQ distributions are the specific ranges where diagnostic questions most often arise (e.g., intellectual disability or giftedness). When examining possible explanations for and sources of IQ differences, the only systematic influence (other than IQ level) was age, as found in Study II. Although several possible influences could not be investigated, the thorough integration of results from both studies allows for making assumptions about the role of between-test variability (e.g., differences in standardization sample) and specific test score characteristics (e.g., number, *g* loadings, and content of subtests included) in test comparability. That is, because in Study II, unlike in Study I, between-test variability was ruled out completely, and transient error was held to a minimum. Thereby, individual-level comparability was indeed remarkably higher in Study II compared to Study I. Yet, individual-level comparability was still unsatisfactory. We therefore hypothesized that characteristics of test scores that systematically vary between FSIQs and ScrIQs (number, *g* loadings, and content of subtests) have influenced the accuracy of scores in a way that reduces individual-level comparability. This is in line with previous research revealing that the most accurate composites are those derived from numerous (at least four), diverse, and highly *g*-loaded subtests (Farmer et al., 2019). However, more research would be needed to estimate the influence of such characteristics on individual-level comparability.

Furthermore, we found that comparability was generally lower among intelligence scores obtained in *different* tests (Study I). This leads to the hypothesis that construct-irrelevant variance between tests might reduce comparability in addition to differences in test score composition. This assumption is supported by the finding from Study I that, in line with the Flynn effect (Flynn, 1987, 2009), IQs were higher for tests with older standardization years. Moreover, it is probable that unsystematic (transient) error reduces comparability among scores obtained in different test batteries.

Finally, both studies concluded that CIs based on overall internal consistency coefficients lead to an overestimation of test score accuracy for diagnostic classification, because a large amount of measurement error is neglected. As demonstrated in Study II, IQ- and age-specific internal consistencies or test–retest reliabilities might better reflect accuracy of intelligence scores when used for individual classification purposes.

In line with previous studies, and unlike for intelligence found in Studies I and II, parent questionnaires and performance-based measures of attention and executive functions did not correlate significantly in Study III. Whereas previous studies have shown that both measurement methods could reveal difficulties in ADHD children (Fried et al., 2015; Gioia et al., 2002), results from our study show that when both measurement methods were administered within the same sample, only questionnaire scores differentiated between children with and without ADHD. Therefore, the two measurement methods were not deemed comparable for group comparisons or for the individual diagnostic setting. There are three possible explanations for this finding: First, the two measurement methods may not fully measure the same underlying construct: Whereas the Conners 3 questionnaires yield two general scores of attention and executive functions but involve questions targeting behavioral and emotional aspects, the CANTAB performance-based tasks differentiate between several subcomponents and exclusively target cognitive processes. Yet, it must be noted that the theoretical background of the Conners 3 as well as other questionnaires remains somewhat

unspecific (Drechsler & Steinhausen, 2013; Lidzba et al., 2013). For instance, the authors of the Conners 3 scales explained that these would measure "the consequences of an inextricable conglomeration of different attention and inhibition processes" (Lidzba et al., 2013, p. 53). Although performance-based measures of attention and executive functions do not consistently rely on the same theoretical framework, like intelligence does often on CHC theory, they often specify targeted components more explicitly (Fray et al., 1996; Grob & Hagmann-von Arx, 2018; Jäger & Horn, 2007).

Second, CANTAB tasks, as for most performance-based measures of attention and executive functions, usually have reliability coefficients that are < .90 and often even < .70 (Soveri et al., 2018; Syväoja et al., 2015) These are considered sufficient to good for research purposes on the group level but rather insufficient for individual diagnostic decisions (Evers, 2001b). This issue might not only (partially) explain the incomparability between tests but also reduce the utility of such performance-based measures for diagnostic purposes. Barkley (2019) even encouraged diagnosticians to refrain from administering performance-based tests for the diagnosis of ADHD, unless their usefulness is proven. However, it may still be essential for diagnosticians to know the children's performance in a standardized test situation in addition to the symptoms reported by parents. This is because of the third possible explanation of incomparability, namely, that questionnaire scores are likely to be biased by the raters' expectations as well as their cultural background (Caye et al., 2017; Thorell et al., 2018). For example, findings from a cross-cultural study showed that differences on the Conners 3 scales were sometimes large enough for a child to be classified as being within the normal range ($t < 60$) in one country and impaired ($t > 70$) in another (Thorell et al., 2018).

Overall, to enhance diagnostic utility of tests measuring attention and executive functions, first a clear explanation of the specific theoretical framework the tests are based on would help practitioners better understand possibly diverging results between the two test methods. Second, and similar to intelligence scores, measurement error should be taken into

account by providing and interpreting CIs around the obtained test scores. Further research should investigate whether measurement error is higher in the vicinity of cut-off scores for tests measuring attention and executive functions, as was found for intelligence tests in Studies I and II. If this is the case, measurement error for individuals who score in these ranges is more likely to lead to classification error (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Our suggestion from Studies I and II to take these circumstances into account by providing conditional reliabilities and standard errors (e.g., age- and IQ-specific coefficients) would then apply to tests measuring attention and executive functions too. Third, future test construction approaches may provide more accurate performance-based tests to overcome described issues (e.g., lack of diagnostic utility due to reduced reliability).

However, independent of what tests are provided from test development and research, test administrators always bear “the ultimate responsibility for appropriate test use and interpretation” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 141). It is important to keep this in mind, because even though tasks and tests can be of high relevance for research focusing on group comparisons, the same tests do not necessarily fulfill the criteria to be applied for individual diagnostic purposes. Therefore, test administrators always need to "know what their tests can do and act accordingly” (Weiner, 1989, p. 829). To take this responsibility, test administrators need to be well trained and informed about the nature and possible limitations of particular psychometric tests. They should rely not only on test manuals but also on actual research and test reviews that provide information about the diagnostic utility and limitations of tests (Diagnostik- und Testkuratorium, 2018). If they are trained well enough, and if test manuals or independent research provides sufficient information about reliability and measurement error, test administrators can calculate the CIs around their obtained scores themselves. This step is important, as the present studies show

that the interpretation of exact test scores obtained from single tests does not hold empirically, which strongly calls into question the use of cut-off scores. Yet, due to restricted reliability, CIs that are specific to age and performance level are quite wide for intelligence (up to 30 IQ points) and can be expected to be so for attention and executive functions scores as well. The extended width of confidence intervals challenges the allocation of a clear diagnosis. One way to deal with this issue may be to conduct at least two tests for the same attribute. If test scores obtained in two different tests are comparable, the interpretation of a test result and subsequent diagnoses and decisions are strengthened. Otherwise, differences in theoretical background, test situations, and individual characteristics that might have differentially interfered with the two tests would need to be considered in order to integrate diverging results in the diagnostic process. Yet, in the longer term, further research might reveal whether and how accuracy of psychometric test scores can be increased in order to also enhance the comparability and diagnostic utility of tests.

One possible approach to enhancing the accuracy of intelligence test scores addresses the circumstance that most contemporary intelligence tests provide composite scores reflecting not only general intelligence but also broad abilities. General intelligence and broad abilities are thereby assessed with the same subtests. As stated by Luecht, Gierl, Tan, and Huff (2006), it is a great challenge to measure something general and something specific with the same instrument. This is because the variance of a set of test scores is limited. As a consequence, if a large amount of variance can be explained by a single aggregated score (e.g., general intelligence score), little variance is left to be unique to any subscore (e.g., factor scores, broad abilities), or vice versa. In this vein, many studies have shown that most factor scores in contemporary intelligence tests contain only a little unique variance compared to the IQ (Canivez & Youngstrom, 2019; McGill, Dombrowski, & Canivez, 2018). Instead of constructing tests that provide general as well as broad ability scores, Luecht et al. (2006) proposed creating multiple unidimensional tests, of which each should be carefully

constructed to measure specific theoretically based cognitive abilities (e.g., fluid intelligence, crystallized intelligence, processing speed, etc.). Transferring findings from Farmer et al. (2019) to this approach, each specific test would need to consist of at least 4 but up to 12 heterogeneous subtests with high specific factor loadings. As found by Beaujean and Benson (2019b), the lower the factor loadings, the more subtests need to be included in a composite score to represent the target attribute as accurately as possible. In turn, if a test is constructed to measure general intelligence, the composite score should consist of (only) subtests with high *g* loadings. This approach could be applicable for tests measuring attention and executive functions as well. This would mean that each component of the theoretical models (e.g., sustained or selective attention; updating, inhibition, and shifting of attention) should be measured with at least four subtests that highly load on the target component and that add up to a composite score. With such a procedure, reliability of a test score may be enhanced and tests will be more useful for individual diagnostic purposes.

However, even if future research supports this approach, there are several challenges associated with it, of which I would like to point out three: First, the construction as well as the administration of such tests would be quite costly and time consuming. Second, there are studies indicating a structural change in the components of executive functions: These studies indicate that the three classic subcomponents of executive functions are relatively undifferentiated and best represented by a single general factor in preschool children, two factors can be differentiated for school-age children, and a three-factor structure is only evident from late childhood on (e.g., Lee, Bull, & Ho, 2013; Roebers, 2017; Wiebe et al., 2011). Such structural change would need to be considered in tests measuring each component separately (ore one single general component) for individual diagnostic purposes. Third, theoretical models of human cognition have developed and changed in the past decades and will most probably continue to do so. For instance, as introduced above, recent theories include attentional processes and executive functions as part of a more comprehensive system

of self-regulation that encompasses biological, emotional, behavioral, and cognitive control and regulation processes (Blair & Raver, 2012, 2015; Calkins & Williford, 2009). Other strands of research include attentional processes and executive functions in an extended version of CHC theory (Jewsbury, Bowden, & Duff, 2017) or in alternative theories of intelligence (e.g., Planning, Attention-Arousal, Simultaneous and Successive [PASS] theory of intelligence; Das, Naglieri, & Kirby, 1994; Naglieri, Das, & Goldstein, 2012), due to their overlap with intelligence (e.g., working memory or processing speed). If future cognitive tests are constructed based on such theories, the composition of a general ability score as well as the more specific abilities to be assessed would be changed, with high expected impact on comparability and diagnostic utility of such tests in school psychology.

Conclusion

Findings from the present research show that standardized scores obtained in different psychometric tests that purport to measure the same psychological cognitive abilities are not necessarily comparable in a way that leads to the same diagnostic classification. Sources of incomparability seem to be differences in or weak theoretical grounding and measurement error. To account for this, the CI and not the exact test score should always be interpreted in tests used for diagnostic purposes. The CI should be calculated on reliability estimates that take measurement error adequately into account. Future research and test construction approaches may provide tests that are even more strongly based on prevailing theories and enhanced in accuracy. Approaches to enhancing the accuracy of diagnostic classification are associated with considerable time and financial costs; yet, given the far-reaching consequences diagnoses may have, it can be considered well worth the effort.

References

Allen, D. N., Stolberg, P. C., Thaler, N. S., Sutton, G., & Mayfield, J. (2014). Validity of the RIAS for assessing children with traumatic brain injury: Sensitivity to TBI and comparability to the WISC-III and WISC-IV. *Applied Neuropsychology: Child*, *3*, 83–93. https://doi.org/10.1080/21622965.2012.700531

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct, including 2010 and 2016 amendments*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*, 65–94. https://doi.org/10.1037/0033-2909.121.1.65

Barkley, R. A. (2019). Neuropsychological testing is not useful in the diagnosis of ADHD: Stop it (or prove it)! *The ADHD Report*, *27*, 1–8. https://doi.org/doi:10.1521/adhd.2019.27.2.1

Baum, K. T., Shear, P. K., Howe, S. R., & Bishop, S. L. (2015). A comparison of WISC-IV and SB-5 intelligence scores in adolescents with autism spectrum disorder. *Autism*, *19*, 736–745. https://doi.org/10.1177/1362361314554920

Beaujean, A. A., & Benson, N. F. (2019a). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. *Applied Measurement in Education*, *32*, 198–215. https://doi.org/10.1080/08957347.2019.1619560

Beaujean, A. A., & Benson, N. F. (2019b). Theoretically-consistent cognitive ability test

development and score interpretation. *Contemporary School Psychology*, *23*, 126–137. https://doi.org/10.1007/s40688-018-0182-1

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology*, *72*, 29–48. https://doi.org/10.1016/j.jsp.2018.12.004

Bergeron, R., Floyd, R. G., & Shands, E. I. (2008). States' eligibility guidelines for mental retardation: An update and consideration of part scores and unreliability of IQs. *Education and Training in Developmental Disabilities*, *43*, 123–131.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*, 327–336. https://doi.org/10.1016/j.lindif.2011.01.007

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*, 31–46. https://doi.org/10.1093/arclin/acn001

Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L' Année Psychologique*, *11*, 191–244. https://doi.org/10.3406/psy.1904.3675

Blair, C., & Raver, C. C. (2012). Individual development and evolution: Experiential canalization of self-regulation. *Developmental Psychology*, *48*, 647–657. https://doi.org/10.1037/a0026472

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*, 711–731. https://doi.org/10.1016/j.physbeh.2017.03.040

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false

belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Calkins, S. D., & Williford, A. P. (2009). Taming the terrible twos: Self-regulation and school readiness. In B. H. Barbarin & O. A. Wasik (Eds.), *Handbook of child development and early education: Research to practice* (pp. 172–198). New York, NY: Guilford Press.

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical, clinical, and policy implications. *Applied Measurement in Education*, *32*, 232–248. https://doi.org/10.1080/08957347.2019.1619562

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. https://doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*, 592.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York, NY: Elsevier.

Caye, A., Machado, J. D., & Rohde, L. A. (2017). Evaluating parental disagreement in ADHD diagnosis: Can we rely on a single report from home? *Journal of Attention Disorders*, *2*, 561–566. https://doi.org/10.1177/1087054713504134

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment ofcognitive processes: The PASS theory of intelligence*. Boston, MA: Allyn and Bacon.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Diagnostik- und Testkuratorium. (2018). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 03. Jan. 2018. *Psychologische Rundschau*, *69*, 109–116.

Diamond, A. (2012). Activities and programs that improve children's executive functions. *Current Directions in Psychological Science*, *21*, 335–341. https://doi.org/10.1177/0963721412453722

Drechsler, R. (2007). Exekutive Funktionen: Übersicht und Taxonomie [Executive functions: Overview and taxonomoy]. *Zeitschrift für Neuropsychologie*, *18*, 233–248. https://doi.org/10.1024/1016-264x.18.3.233

Drechsler, R., & Steinhausen, H. C. (2013). *Verhaltensinventar zur Beurteilung exekutiver Funktionen BRIEF. Deutschsprachige Adaption des Behavior Rating Inventory of Executive Function*. Bern, Switzerland: Hogrefe.

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, *1*, 137–153. https://doi.org/10.1207/S15327574IJT0102

Evers, A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing*, *1*, 155–182. https://doi.org/10.1207/S15327574IJT0102

Farmer, R. L., Floyd, R. G., Reynolds, M. R., & Berlin, K. S. (2019). How can general intelligence composites most accurately index psychometric g and what might be good enough ? *Contemporary School Psychology*. https://doi.org/10.1007/s40688-019-00244-1

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice*, *39*, 414–423. https://doi.org/10.1037/0735-7028.39.4.414

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 427–427. https://doi.org/10.1037/h0090408

Flynn, J. R. (2009). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938-2008. *Economics and Human Biology*, *7*, 18–27. https://doi.org/10.1016/j.ehb.2009.01.009

Fray, P., Robbins, T. W., & Sahakian, B. J. (1996). Neuropsychiatric applications of CANTAB. *International Journal of Geriatric Psychiatry*, *11*, 329–336. https://doi.org/https://doi.org/10.1002/(SICI)1099-1166(199604)11:4%3C329::AID-

GPS453%3E3.0.CO;2-6

Fried, R., Hirshfeld-Becker, D., Petty, C., Batchelder, H., & Biederman, J. (2015). How informative is the CANTAB to assess executive functioning in children with ADHD? A controlled study. *Journal of Attention Disorders*, *20*, 1–8. https://doi.org/10.1177/1087054712457038

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science Quarterly*, *50*, 21–43.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Test review: Behavior Rating Inventory of Executive Function. *Child Neuropsychology : A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *6*, 235–238. https://doi.org/10.1076/chin.6.3.235.3152

Gioia, G. A., Isquith, P. K., Kenworthy, L., & Barton, R. M. (2002). Profiles of everyday executive function in acquired and developmental disorders. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *8*(2), 121–137. https://doi.org/10.1076/chin.8.2.121.8727

Goldstein, S., Princiotta, D., & Naglieri, J. A. (Eds.). (2015). *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts*. https://doi.org/10.1007/978-1-4939-1562-0

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

Grob, A., Gygi, J. T., & Hagmann-von Arx, P. (2019). *The Stanford-Binet Intelligence Scales, Fifth Edition (SB5)—German Adaptation*. Bern, Switzerland: Hogrefe.

Grob, A., & Hagmann-von Arx, P. (2018). *Intelligence and Development Scales–2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche [Intelligence and Development Scales for children and adolescents]*. Bern, Switzerland: Hogrefe.

Hagmann-von Arx, P., & Grob, A. (2014). *Reynolds Intellectual Assessment Scales (RIAS)—German adaptation*. Bern, Switzerland: Hans Huber.

Hagmann-von Arx, P., Lemola, S., & Grob, A. (2018). Does IQ = IQ ? Comparability of intelligence test scores in typically developing children. *Assessment*, *25*, 691–701. https://doi.org/10.1177/1073191116662911

Heyanka, D. J., Holster, J. L., & Golden, C. J. (2013). Intraindividual neuropsychological test variability in healthy individuals with high average intelligence and educational attainment. *International Journal of Neuroscience*, *123*, 526–531. https://doi.org/10.3109/00207454.2013.771261

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*, 253–270. https://doi.org/10.1037/ h0023816

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*, 93–114.

Jäger, R. S., & Horn, R. (2007). *TEA-Ch—Test of Everyday Attention for Children. German adaptation*. Frankfurt am Main, Germany: Pearson Assessment.

Jensen, A. R., & Weng, L.-J. (1994). What is a good *g*? *Intelligence*, *258*, 231–258.

Jewsbury, P. A., Bowden, S. C., & Duff, K. (2017). The Cattell–Horn–Carroll model of cognition for clinical assessment. *Journal of Psychoeducational Assessment*, *35*, 547–567. https://doi.org/10.1177/0734282916651360

Katsuki, F., & Constantinidis, C. (2014). Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, *20*, 509–521. https://doi.org/10.1177/1073858413514136

Kovacs, K., & Conway, A. R. A. (2019). A unified cognitive/differential approach to human intelligence: Implications for IQ testing. *Journal of Applied Research in Memory and Cognition*, *8*, 255–272. https://doi.org/10.1016/j.jarmac.2019.05.003

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, *84*, 1933–1953. https://doi.org/10.1111/cdev.12096

Lidzba, K., Christiansen, H., & Drechsler, R. (2013). *Conners Skalen zu Aufmerksamkeit und Verhalten-3. Deutschsprachige Adaptation der Conners 3rd edition von Keith Conners*. Bern, Switzerland: Hogrefe.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006). Scalability and the development of useful diagnostic scales. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.*

Maedgen, J. W., & Carlson, C. L. (2000). Social functioning and emotional regulation in the attention deficit hyperactivity disorder subtypes. *Journal of Clinical Child Psychology*, *29*, 30–42. https://doi.org/10.1207/S15374424jccp2901

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment*, *26*, 207–214. https://doi.org/10.1037/a0034832

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. https://doi.org/10.1016/j.jsp.2018.10.007

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–181). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–182). New York, NY: Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the

shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1–10.

Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review*, *2*, 109–145. https://doi.org/10.1007/BF01109051

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., … Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 2693–2698. https://doi.org/10.1073/pnas.1010076108

Naglieri, J. A., Das, J. P., & Goldstein, S. (2012). Planning, attention, simultaneous, successive: A cognitive-processing-based theory of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 178–194). New York, NY: Guilford Press.

National Research Council. (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academy Press.

Neukrug, E., & Fawcett, C. (2014). *Essentials of testing and assessment: A practical guide for counselors, social workers, and psychologists.* Belmont, CA: Thomson Brooks/Cole.

Petermann, F. (2009). *Wechsler Preschool and Primary Scale of Intelligence–Third Edition (WPPSI-III; German Adaptation)*. Frankfurt am Main, Germany: Pearson Assessment.

Petermann, F., & Petermann, U. (2011). *Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; German adaptation)*. Frankfurt am Main, Germany: Pearson Assessment.

Polanczyk, G. V, Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD

prevalence estimates across three decades: An updated systematic review and meta-regression analysis. *International Journal of Epidemiology*, *43*, 434–442. https://doi.org/10.1093/ije/dyt261

Posner, M. I., & Rothbart, M. K. (2007). Research on attention networks as a model for the integration of psychological science. *Annual Review of Psychology*, *58*, 1–23. https://doi.org/10.1146/annurev.psych.58.110405.085516

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Roebers, C. M., Röthlisberger, M., Cimeli, P., Michel, E., & Neuenschwander, R. (2011). School enrolment and executive functioning: A longitudinal perspective on developmental changes, the influence of learning context, and the prediction of pre-academic skills. *European Journal of Developmental Psychology*, *8*, 526–540. https://doi.org/10.1080/17405629.2011.571841

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206–224. https://doi.org/10.1037/1082-989X.8.2.206

Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2014). Measurement theory in action: Case studies and exercises. In *Measurement Theory in Action: Case Studies and Exercises* (2nd ed.). https://doi.org/10.4135/9781452224749

Soveri, A., Lehtonen, M., Karlsson, L. C., Lukasik, K., & Antfolk, J. (2018). Test–retest reliability of five frequently used executive tasks in healthy adults. *Applied*

Neuropsychology:, *25*, 155–165. https://doi.org/10.1080/23279095.2016.1263795

Spearman, C. E. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Blackburn Press.

Stern, E., & Neubauer, A. (2016). Intelligenz: kein Mythos , sondern Realität [Intelligence: Not a myth but reality]. *Psychologische Rundschau*, *67*, 1–13. https://doi.org/10.1026/0033-3042/a000290

Swiss Conference of Cantonal Ministers of Education. (2007). Interkantonale Vereinbarung über die Zusammenarbeit im Bereich der Sonderpädagogik [Intercantonal agreement on collaboration in special education]. Retrieved from http://www.edk.ch

Syväoja, H. J., Tammelin, T. H., Ahonen, T., Räsänen, P., Tolvanen, A., Kankaanpää, A., & Kantomaa, M. T. (2015). Internal consistency and stability of the CANTAB neuropsychological test battery in children. *Psychological Assessment*, *27*, 698–709. https://doi.org/10.1037/a0038485

Tellegen, P. J., Laros, J. A., & Petermann, F. (2012). *SON-R 6-40. Snijders-Oomen Nonverbaler Intelligenztest [SON-R 6-40 Snijders-Oomen Nonverbal Intelligence Test]*. Göttingen, Germany: Hogrefe.

Thorell, L. B., Chistiansen, H., Hammar, M., Berggren, S., Zander, E., & Bölte, S. (2018). Standardization and cross-cultural comparisons of the Swedish Conners 3 rating scales. *Nordic Journal of Psychiatry*, *72*, 613–620. https://doi.org/10.1080/08039488.2018.1513067

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry*, *54*, 131–143. https://doi.org/10.1111/jcpp.12001

van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2013). The structure of executive functions in children: A closer examination of inhibition, shifting, and updating. *British Journal of Developmental Psychology*, *31*, 70–87.

https://doi.org/10.1111/j.2044-835X.2012.02079.x

van Zomeren, A. H., & Brouwer, W. H. (1994). *Clinical neuropsychology of attention*. Oxford University Pres.

von Aster, M. G., Neubauer, A., & Horn, R. (2006). *Wechsler Intelligenztest für Erwachsene (WIE-III) [German adaptation of the Wechsler Adult Intelligence Scale III]*. Frankfurt am Main, Germany: Harcourt Test Services.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment*, *53*, 827–831. https://doi.org/doi.org/10.1207/s15327752jpa5304_18

Weissman, H. N., & Debow, D. M. (2003). Ethical principles and professional competencies. In A. M. Goldstein (Ed.), *Handbook of psychology: Forensic psychology* (pp. 33–53). Hoboken , NJ: Wiley & Sons.

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A. C., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*, 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V, & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry*, *57*, 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006

World Health Organization [WHO]. (1990). *The International Statistical Classification of Diseases and Related Health Problems* (10th ed.). Geneva, Switzerland: Author.

World Health Organization [WHO]. (2018). *International Classification of Diseases for Mortality and Morbidity Statistics* (11th rev.). Retrieved from https://icd.who.int/browse11/l-m/en

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and

attention. *Monographs of the Society for Research in Child Development*, *78*, 16–33. https://doi.org/10.1111/mono.12032

APPENDIX A: Study I

Bünger, A., Grieder, S., Schweizer, F., & Grob, A. (2019). The comparability of intelligence test results: Group- and individual-level comparisons of seven intelligence tests. *Manuscript submitted to the Journal of School Psychology.*

Draft January 20, 2020

The Comparability of Intelligence Test Results: Group- and Individual-Level Comparisons of Seven Intelligence Tests

Anette Bünger, Silvia Grieder, Florine Schweizer, and Alexander Grob

University of Basel

Author Note

Anette Bünger, Silvia Grieder, Florine Schweizer, and Alexander Grob, Department of Psychology, Division of Developmental and Personality Psychology, University of Basel.

Funding: This work was supported by the Suzanne and Hans Biaesch Foundation for the Enhancement of Applied Psychology.

Declarations of interest: none.

Correspondence concerning this article should be addressed to Anette Bünger, Department of Psychology, Division of Developmental and Personality Psychology, University of Basel, Missionsstrasse 62, 4055 Basel, Switzerland, email: anette.buenger@unibas.ch

Abstract

A large body of research shows that IQs obtained from different intelligence tests substantially correlate on the group level. Yet, there is little research investigating whether different intelligence tests yield comparable results for individuals. This is, however, paramount to the application of intelligence tests in practice, as high-stakes decisions are based on individual test results. We therefore investigated whether seven current and widely used intelligence tests yield comparable results for individuals aged 4–20 years. We found mostly substantial correlations, though several significant mean differences on the group level. Results on individual-level comparability indicate that the interpretation of exact intelligence test scores does not hold. Even the 95% confidence intervals could not be reliably replicated with different intelligence tests. Further, the nominal level of intelligence systematically predicted IQ differences between tests. In this vein, individual-level comparability was lower in the below-average and above-average intelligence ranges compared to the average range, even after considering the regression toward the mean. Yet these are the specific ranges in which diagnostic questions most often arise in practice. Similar patterns appeared for the individual-level comparability of nonverbal and verbal intelligence factor scores as well. Implications for test interpretation and test construction are discussed.

Keywords

intelligence tests, IQ, comparability, individual-level, children and adolescents

In the field of applied educational, learning, and school psychology, psychometric test batteries are often applied to assess the developmental status and abilities of a child or adolescent. These test results provide the basis for far-reaching decisions concerning educational support, interventions, and therapy (e.g., curative education, special education programs, support for gifted children, attention training). When making an educational decision, one of the most frequently assessed domains is an individual's cognitive ability, measured by intelligence tests. This might be because many diagnoses that are relevant in the field of school psychology can only be given in relation to a standardized intelligence test, as recommended by the current versions of the *International Statistical Classification of Diseases and Related Health Problems* (World Health Organization, 2018) and the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013). In addition, intelligence is one of the most investigated constructs of psychology and numerous tests are available in many languages to provide age-standardized intelligence scores, that is, IQs (Lubinski, 2004). Furthermore, numerous studies have confirmed the association between intelligence and central areas of life, such as educational success (Deary, Strand, Smith, & Fernandes, 2007), professional success and income (Damian, Su, Shanahan, Trautwein, & Roberts, 2015), success in interpersonal relationships (Aspara, Wittkowski, & Luo, 2018), health (Wrulich et al., 2014), and even longevity (Calvin et al., 2011).

Group- and Individual-Level Comparability

Currently, a large number of different intelligence tests are available to diagnosticians. For example, according to a keyword search, there are 39 intelligence tests available for purchase via *Testzentrale* (http://www.testzentrale.ch), an official distributor of German psychometric tests. As German tests are often adaptations of U.S. originals, there are probably even more widely used intelligence tests available in the United States. The opportunity to choose between different intelligence tests has several advantages. For example, learning effects can be minimized by selecting an alternative test battery when testing multiple times.

Moreover, additional test batteries can be used to confirm the results of one test. Finally, special characteristics of an individual or of a specific counseling situation can be taken into account by, for instance, providing nonnative speakers with tests that are not specific to any one culture (Hagmann-von Arx, Petermann, & Grob, 2013), or by using a short test (screening) when intelligence is assessed as a control variable (Hagmann-von Arx & Grob, 2014). Diagnosticians are advised to exclusively apply intelligence tests with strong evidence for reliability and validity (International Test Commission, 2001) in terms of content, construct (i.e., convergent and divergent), and criterion (i.e., concurrent and predictive power) validity. Tests that meet these criteria are assumed to render reliable and valid results, and IQs obtained in different test batteries are expected to be comparable (Floyd, Clark, & Shadish, 2008) if they are composed of multiple, diverse, and reliable subtests with high loadings on a general intelligence (*g*) factor (Jensen & Weng, 1994). This assumption is supported by the principle of aggregation (Rushton, Brainerd, & Pressley, 1983), which states that the sum of multiple measurements (IQ based on multiple subtests) represents a more stable predictor than a single measurement (IQ based on a single subtest) because measurement error is averaged out. Moreover, a large body of research has confirmed that many intelligence test batteries are highly correlated (Allen, Stolberg, Thaler, Sutton, & Mayfield, 2014; Baum, Shear, Howe, & Bishop, 2015; Hagmann-von Arx, Grob, Petermann, & Daseking, 2012; Hagmann-von Arx, Lemola, & Grob, 2018), which can be seen as an indication of convergent validity, that is, that the test batteries measure the same constructs (Neukrug & Fawcett, 2014). These results are based on group-level comparisons.

Yet, the fact that different intelligence tests correlate highly on the group level is a necessary but not a sufficient criterion for comparability, because diagnoses are given to individuals and because decisions are made based on individual intelligence test results. To ensure that these decisions and diagnoses are valid, the tests must also be comparable on the individual level. This means that it should not matter which test is used to assess an

individual's intelligence, as the IQ should always be equal given a certain error tolerance. Floyd et al. (2008) referred to this type of comparability as the *exchangeability of IQs*. As yet, there is surprisingly little research on this, with only two studies available that investigated the comparability of IQs on the individual level. Both studies used the overlap of the 90% confidence intervals (CIs) of each intelligence test pair as criterion for comparability (Floyd et al., 2008; Hagmann-von Arx et al., 2018). The advantage of using CIs is that it takes the unreliability of measurements into account. However, when using CIs, the definition of comparability varies across test comparisons, as reliability and CIs differ between tests. Floyd et al. (2008) therefore used a second criterion for comparability—an absolute difference between two scores less than or equal to 10 IQ points—which is in line with guidelines that recommend considering 5 IQ points above and below the obtained IQ when diagnosing mental retardation (i.e., Bergeron, Floyd, & Shands, 2008). In the first study, Floyd et al. (2008) compared seven English intelligence tests with standardization years between 1987 and 2003 for an age range of 8 to 16 years, with one sample also covering undergraduate students. Findings showed that the IQs differed in about 36% of intelligence test pairs when considering the 90% CI, and in about 28% when considering the 10-point IQ criterion. In the second study, Hagmann-von Arx et al. (2018) compared five German intelligence tests for 6- to 11-year-old children and with standardization years between 2003 and 2012. The authors found that IQs differed in about 24% of intelligence test pairs when considering the 90% CI.

The two studies indicate that for 64% to 76% of schoolchildren, the IQs obtained from different tests can be considered comparable on the individual level. This means that decisions and diagnoses based on a single test are of questionable validity for a fourth to a third of children.

Reasons for (In-)Comparability

To better understand this relatively high percentage of incomparability, it is important to consider potential sources of variability in intelligence test scores that do not stem from

true ability of the examinees. In addition to unsystematic random errors, these sources include specific test characteristics, characteristics of the test situation (environmental influences), personal characteristics of the examinees, and examiner's influence (National Research Council, 2002).

Test batteries that are applied for important decisions must provide reliable test results, that is, have reliability coefficients of at least .90 (Evers, 2001). This means that even if the composite score of a particular test fulfills this criterion, it always contains an unsystematic random error component. This is usually taken into account with the CI plotted around the obtained IQ. Furthermore, it is possible that test batteries lead to inconsistent test results because of differences in standardization. For example, the historical time of standardization may be an influential aspect, as demonstrated by Flynn (1987, 2009). The Flynn effect refers to the fact that the average intelligence of a population increases between 3 and 5 IQ points per decade. In contrast, since the end of the 1990s, there have also been studies showing a negative or so-called anti-Flynn effect. Authors of these studies postulate that average IQs have been declining again since the 1990s (Dutton & Lynn, 2014; Lynn & Harvey, 2008; Sundet, Barlaug, & Torjussen, 2004). Regardless of whether there is an increase or decrease in average intelligence, performance on intelligence tests appears to be influenced by culture and cohort effects. This issue can play an important role in the comparability of test results, especially if the standardization years of two tests are far apart. Moreover, IQs are not a pure representation of an individual's true cognitive ability (intelligence construct or true *g*). Therefore, it is important to consider test characteristics that affect measurement accuracy and that vary across intelligence tests (Farmer, Floyd, Reynolds, & Berlin, 2019). For instance, most contemporary intelligences tests relate to the Cattel–Horn–Carroll (CHC) theory of cognitive abilities (McGrew, 1997, 2005, 2009). Yet, the composite score of an intelligence test, which is considered an indicator of *g*, may be calculated based on different sets of subtests that vary in number as well as heterogeneity of content. Especially for IQs that are

based on only a few subtests, inadequate content sampling poses a risk for inflation or deflation of IQs and may therefore reduce their comparability (Farmer et al., 2019). In addition, systematic differences in task format may occur (e.g., verbal or nonverbal, paper-and-pencil or computer based Floyd et al., 2008). This may interact with conditions and characteristics of individual examinees such as limited language skills, for example, due to migration background (Daseking, Lipsius, Petermann, & Waldmann, 2008), or any type of speech or language disorder (Gallinat & Spaulding, 2014; Miller & Gilbert, 2008), as well as impaired fine motor skills (Yin Foo, Guppy, & Johnston, 2013). Regarding characteristics of a specific test situation, the order of test administration with practice or fatigue effects (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007), the time interval between two or more tests, or the time of day (Gupta, 1991) in combination with "morningness" and "eveningness" of examinees (Goldstein, Hahn, Hasher, Wiprzycka, & Zelazo, 2007) as well as general motivation in the specific test situation (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011) may cause variance in intelligence test scores that does not reflect an examinee's true ability. To investigate to what extent intelligence test scores are influenced by individuals and test procedures, Hagmann-von Arx et al. (2018) and Floyd et al. (2008) applied generalizability theory analyses (Briesch, Swaminathan, Welsh, & Chafouleas, 2014; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Both studies showed that no more than 4% of the variance in IQs could be attributed to specific test characteristics. The largest part of the variance (7%–27%, Floyd et al., 2008; 29%–42%, Hagmann-von Arx et al., 2018) was attributable to interactions between the examinees and the test situation. However, only Hagmann-von Arx et al. (2018) explored the impact of specific variables, such as time interval between two tests, the order of test administration, or qualitative nominal intelligence levels, on intelligence test differences and they could not detect any significant systematic effect.

Limitations of Previous Studies

The two previous studies that focused on individual-level comparability of intelligence tests results show several limitations. First, they did not cover the whole age span in which questions relevant to school psychology tend to arise. Furthermore, age was not investigated as a possible predictor of test score differences, and although Hagmann-von Arx et al. (2018) discussed language skills and migration background as a possible influence on test comparability, they did not include any language or cultural background variable in their analyses. Moreover, to investigate the possible impact of intelligence level on test comparability, Hagmann-von Arx et al. (2018) calculated a mean IQ across all intelligence tests that they carried out. This practice is problematic as it obfuscates the influence of systematic errors on the specific test results (Schneider & McGrew, 2011). Another limitation is that both studies investigated the overlap of the 90% CIs, whereas in practice it is most often the 95% CI that is used for interpretation. Last, previous studies investigated only individual-level comparability for general intelligence composites scores (i.e., the full-scale IQ, hereafter referred to simply as IQ) but not for factor index scores, such as verbal and nonverbal factor indexes. Factor index scores are used to identify specific strengths and difficulties (cognitive profile analysis) by about 49% to 55% of school psychologists in the United States (Benson, Floyd, Kranzler, Eckert, & Fefer, 2018), despite growing evidence against their usefulness (McGill, Dombrowski, & Canivez, 2018). Therefore, it seems essential to shed light on the individual-level comparability for factor indexes.

The Present Study

The aim of our study was to gain insight into the comparability of intelligence test scores with a main focus on individual-level comparability, addressing the aforementioned research gaps and limitations from previous studies. For this purpose, we assessed seven current and widely used intelligence tests across childhood and adolescence in German-speaking countries. The wide age range of participants in our sample (4–20 years) covers the preschool, primary, and secondary school years as well as the years after high school

graduation, representing the entire age range of development in which school-psychology issues arise. In accordance with previous research, we expected substantial correlations (Hypothesis 1) and comparable mean scores (Hypothesis 2) on the group level. In addition, we explored the comparability of IQs on an individual level (Research Question 1) using several criteria for comparability including criteria that consider reliability (i.e., overlap of the 90% and 95% CIs) as well as criteria that are independent of reliability and therefore invariant across test batteries (i.e., maximum absolute difference of 10 IQ points and nominal intelligence level). We also explored whether the year of standardization (with respect to a possible Flynn or anti-Flynn effect) influenced comparability such that tests with earlier standardization years revealed higher IQs than tests with a more recent standardization or vice versa (Research Question 2). Moreover, we investigated the individual-level comparability for nonverbal (nonverbal intelligence; NI) and verbal (verbal intelligence; VI) factor indexes of each intelligence test and examined whether individual-level comparability was significantly higher or lower for NI and VI versus IQ (Research Question 3). Last, we investigated whether age, bilingualism, time interval between two tests, order of test administration, and nominal intelligence level were significantly related to IQ differences as an indicator of comparability (Research Question 4). We think providing precise evidence on these questions will have vast consequences for theory and practice in intelligence testing.

Method

Participants

The sample was a subsample of the standardization and validation study of the Intelligence and Development Scales–2 (IDS-2; Grob & Hagmann-von Arx, 2018) as well as of the German adaptation of the Stanford–Binet Intelligence Scales–Fifth Edition (SB5; Grob, Gygi, & Hagmann-von Arx, 2019). We included 383 children and adolescents (M_{age} = 10.31 years, SD_{age} = 3.96; age range: 4 to 20 years; 47% boys). The percentage of bilingual and nonnative speakers was 32%. All participants spoke German well enough to understand

instructions and to give oral answers. None of the participants had a speech or language disorder. According to parent and self-reports, 10 of the included participants were diagnosed with a learning disorder (e.g., dyslexia or dyscalculia), one was diagnosed with Asperger's syndrome, and 15 were diagnosed with attention-deficit/hyperactivity disorder. The pattern of test comparability did not change with the inclusion of these participants; hence they were not excluded from the study. Forty-eight percent of the participants' mothers held a university degree.

Measures

For the intelligence test comparisons, we applied the IDS-2 (Grob & Hagmann-von Arx, 2018), the German adaptations of the SB5 (Grob et al., 2019), the Reynolds Intellectual Assessment Scales (RIAS; Hagmann-von Arx & Grob, 2014), the Snijders Oomen Nonverbal Intelligence Test 6-40 (SON-R 6-40; Tellegen, Laros, & Petermann, 2012), the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III; von Aster, Neubauer, & Horn, 2006), the Wechsler Intelligence Scales for Children–Fourth Edition (WISC-IV; Petermann & Petermann, 2011), and the Wechsler Preschool and Primary Scale of Intelligence–Third Edition (WPPSI-III; Petermann, 2009). Details on the intelligence tests such as age range, latent factors measured, year of standardization, and reliability (internal consistencies), as well as indications of which latent factors were classified as IQ, NI, and VI, can be found in Table 1.

Procedure

Recruitment was carried out for preschoolers via day-care centers, nursery schools, and kindergartens and for school-aged children and adolescents via schools, institutions, and private contacts. All participants were assessed with the IDS-2 and/or the SB5 and with one or more of the above-mentioned intelligence tests that were suitable for their age (as described in Table 1). Due to restrictions in testing time, some participants could not be assessed with all measures that were suitable for their age. The assessments were conducted between June 2015

and February 2018 in individual settings by trained psychology major students. The tests took place at participants' homes or at the University of Basel. The participants (16- to 20-year-olds) or their parents (4- to 15-year-olds) received written feedback on the test results, where standard scores were available (i.e., RIAS, SON-R 6-40, WAIS-III, WISC-IV, WPPSI-III). For the standardization study, participants received a voucher worth CHF 30 (for a 3 to 3.5-h test; i.e., IDS-2) or CHF 20 (for a 2-h test; i.e., SB5) instead of written feedback, as standard scores were not available for these tests at that time. The Ethics Committee of Northwest and Central Switzerland approved the study protocol. Parents gave written informed consent for their child to participate and participants gave oral (children) or written (adolescents) informed assent.

Statistical Analyses

In a preliminary step, the raw test scores of the intelligence tests were transformed into the respective IQs ($M = 100$, $SD = 15$) using the test manuals. The assumption of normal distribution was tested using the Shapiro–Wilk test. The IQs of the IDS-2, SB5, RIAS, SON-R 6-40, WAIS-III, and WPPSI-I were normally distributed ($W = .959–.994$, $p > .05$). The distribution of the WISC-IV scores deviated significantly from a normal distribution ($W = .967$, $p < .001$), and therefore nonparametric methods were used for the intelligence test comparisons involving the WISC-IV. We tested whether the mean IQs in our sample differed from the expected mean of 100 using single-sample *t* tests and a single-sample Wilcoxon signed-rank test for the WISC-IV. The analyses described in the following refer to the comparison of all intelligence test pairs, except the Wechsler scales (i.e., WAIS-III, WISC-IV, and WPPSI-II). They were not compared among each other, since these each apply to different age ranges. Pearson's product-moment correlations and Spearman's rank-order correlations were calculated for all intelligence test pairs and were further corrected for unreliability (Hypothesis 1).

To test for group-level comparability of the intelligence test scores, *t* tests for dependent samples or Wilcoxon signed-rank tests were calculated (Hypothesis 2). As effect sizes, Cohen's *d* was calculated for *t* tests and *r* for Wilcoxon sign-rank tests. Group-level differences were interpreted as substantial if statistical significance was given at $p < .05$ and an effect of $d \geq 0.20$ or $r \geq .10$ was reached.

To test for the comparability of IQs at the individual level (Research Question 1), we applied five different criteria to define comparability. First, we calculated the sum of half of the 90% CIs of two tests as the critical value, following Floyd et al. (2008). Two IQs of the same individual were considered comparable if the difference between the IQs was smaller than this value (Criterion CI90), which means that the 90% CIs of the two test scores overlapped. Since in practice it is often the 95% CI and not the 90% CI that is interpreted, the overlap of the 95% CIs of two test scores of an individual were also investigated as a second indicator of comparability (Criterion CI95).

To be able to compare our results to findings from previous studies (Floyd et al., 2008; Hagmann-von Arx et al., 2018), we plotted the 90% CIs around the obtained test scores using the standard error of the mean. However, following recommendations of Atkinson (1989) and to correct for the regression toward the mean, most intelligence tests use the estimated true score instead of the obtained test score as well as the standard error of estimation to plot the CI, as is the case for the seven intelligence tests we investigated in this study. Therefore, we adhered to this procedure for the calculation of 95% CIs. As a third criterion, following Floyd et al. (2008), we assumed comparability if the absolute difference between two IQs of an individual was less than or equal to 10 points (Criterion IQ10). Finally, the correspondence of the nominal intelligence level (according to Grob, Reiman, Gut, & Frischknecht, 2013: lower extreme IQ: < 70; below average IQ: 70–84; average IQ: 85–115; above average IQ: 116–130; upper extreme IQ: 130; Criterion Nominal IQ) as well as the correspondence of the

nominal intelligence levels for the 95% CIs (e.g., average to above average; Criterion Nominal CI95) was calculated.

To investigate if there was an effect of standardization time (with respect to a possible Flynn or anti-Flynn effect), we calculated the percentage of participants who scored higher on the more recently standardized intelligence test and the percentage of participants who scored lower on the more recently standardized intelligence test at Criterion CI90 for each pair of intelligence tests (Research Question 2). To examine the individual-level comparability of NI and VI (Research Question 3), we used the CI95 as a criterion that considers test reliability as well as the IQ10 as a criterion that is independent of test reliability and therefore invariant across all test batteries. We then conducted χ^2 tests to examine whether individual-level comparability was significantly higher or lower in NI and VI versus IQ. Finally, to analyze whether age, bilingualism, time interval between two tests, order of test administration, and nominal intelligence level significantly predicted the comparability of test results, we ran simple linear regressions with the described variables as independent variables and absolute IQ differences between each test pair as separate dependent variables (Research Question 4). Age and time interval between two tests were included as continuous variables and bilingualism and order of test administration as binary variables. In contrast to the Criterion Nominal IQ (as described above), nominal intelligence level was defined here as a three-level categorical variable (i.e., below average IQ: < 85; average IQ: 85–115; above average IQ: > 115) due to small sample sizes at the tails. For each test comparison, a participant was classified as below average or above average if at least one of the two intelligence test scores fell within the defined range.

Results

Preliminary Analyses

Table 2 shows the distribution of the intelligence test scores collected in our study. Single-sample *t* tests revealed that means of all IQs (M = 102.46–115.29, p = .007 to p < .001)

were significantly higher than the expected mean (M = 100). Further, the standard deviations of all intelligence tests (SD = 9.64–12.80) except for the SON-R 6-40 (SD = 15.17) were smaller compared to the expected standard deviation (SD = 15).

Group-Level Comparisons

Table 3 shows the results of all 18 intelligence test comparisons on the group level. The first of each pair of intelligence tests is the one with the more recent standardization year, except for the IDS-2 and the SB5, which were standardized within the same time period. Due to the restricted standard deviations reported above, Pearson's product-moment correlation coefficients and Spearman's rank-order correlation coefficients were corrected not only for attenuation but also for range restriction, according to the formula proposed by Alexander, Carson, Alliger, and Carr (1987, p. 312). Sixteen of 18 correlations were significant and ranged from moderate to strong. Only the correlations between the SB5 and the WAIS-III (r = .38, p = .25) and between the SON-R 6-40 and the WAIS-III (r = .37, p = .08) were rather weak and did not reach statistical significance. Further, t tests for dependent samples and Wilcoxon tests showed that the means at the group level were substantially lower for the first-listed test with more recent standardization years in 13 of the 18 comparisons (M_{diff} = 2.81 to -13.43, d = -0.27 to -1.29, r = -.23 to -.81). Only regarding the comparisons between the SON-R 6-40 and the WPPSI-III (M_{diff} = 9.28, d = 0.73) was the mean score higher for the SON-R 6-40 with more recent standardization years. The comparisons between the IDS-2 and the SB5, between the SB5 and the RIAS, between the SB5 and the WAIS-III, and between the RIAS and the WPPSI-III did not reveal substantial differences.

Individual-Level Comparisons

Table 4 shows the results of the intelligence test comparisons on the individual level (Research Question 1). A total of 1,774 test comparisons were analyzed across all seven intelligence tests and 383 subjects. Since the number of possible comparisons between the different test pairs varied considerably, the total percentage of comparability was calculated

across all 1,774 test comparisons. The total comparability regarding the Criteria CI90, IQ10, CI95, Nominal IQ, and Nominal CI95 was 49.8%, 58.1%, 61.9%, 61.9%, and 89.4%, respectively. Regarding Research Question 2, in 39.9% of all cases a lower IQ was achieved in the more recent (first-mentioned) test and in only 9.6% of all cases was a higher IQ achieved in the more recent test. The comparison between the IDS-2 and the SB5 was excluded from this analysis as these tests were standardized within the same time period.

Detailed results on the individual-level comparability of NI and VI (Research Question 3) are reported in Supplemental Tables 1 and 2. For NI, a total of 1,780 test comparisons were analyzed across all seven test batteries, including the IQ of the SON-R 6-40, as it is officially classified as NI. For VI, a total of 1,020 test comparisons were analyzed across all tests except the SON-R 6-40, as this test does not provide any verbal index score. Regarding Criterion IQ10, χ^2 tests revealed that with 49.3%, the total comparability of NI was significantly lower than the total comparability of IQ, with 58.1% ($\chi^2 = 27.62$, $p < .001$) and the total comparability of VI with 62.2% was significantly higher than the total comparability of IQ ($\chi^2 = 6.02$, $p < .05$). When considering Criterion CI95, the total comparability of NI increased to 62.2% and did not differ significantly from the total comparability of IQ, with 61.9% ($\chi^2 = 0.03$, $p > .05$). The total comparability of VI also increased to 68.6% and was significantly higher than the total comparability of IQ ($\chi^2 = 12.60$, $p < .001$).

Predictors of Incomparability

Results of linear regressions with age, bilingualism, time interval between two tests, order of test administration, and nominal intelligence level as predictors of IQ differences (Research Question 4) are reported in Table 5. Age was a significant positive predictor (higher IQ differences for older participants) for only 1 of 18 IQ differences, and a significant negative predictor (lower IQ differences for older participants) for 3 of 18 test comparisons. Bilingualism was a significant positive predictor (higher IQ differences for bilingual compared to monolingual participants) for three test comparisons. Time interval between tests

was a positive predictor (higher IQ differences with larger time interval between two tests) in one test comparison. Order of test administration was a significant positive predictor for two test comparisons.

Regarding nominal intelligence level, having an IQ in the below-average range in at least one intelligence test was significantly associated with higher IQ differences in six test comparisons. Two of the 18 analyses (SB5 vs. WAIS-III; SON-R 6-40 vs. WPPSI-III) could not be completed as none of the participants who completed either of the test combinations achieved an IQ in the below-average range. Having an IQ in the above-average range in at least one intelligence test was significantly associated with higher IQ differences in 15 test comparisons ($\beta = .18$ to .70 and all $p \leq .018$) and significantly associated with lower IQ differences in one test comparison.

Post Hoc Analyses

The systematic effect of the nominal intelligence level on IQ differences revealed in the regression analyses, together with the fact that the mean IQs in our sample were significantly higher than the standardization sample means in all intelligence tests, may have reduced the overall comparability. It therefore may have limited the generalizability of our findings. To address this issue, we additionally ran post hoc analyses on the individual-level comparability for the three nominal intelligence levels (below average: IQ < 85; average IQ: 85–115; above average IQ: > 115) separately. In the below-average IQ group, a total of 144 intelligence test comparisons could be analyzed. In the average IQ group, the total of possible comparisons amounted to 608, and in the above-average IQ group, it was 972. The total percentages of comparability of the three groups are shown in Table 4. Regarding the Criterion CI90, the comparability increased to 61.4% in the average group and decreased to 44.4% each in the above- and below-average groups, compared to the overall comparability. Regarding Criterion IQ10, the comparability increased to 71.2% in the average group and decreased to 50.7% in the below-average and to 51.5% in the above-average group.

Regarding the Criterion CI95, the comparability increased to 75.6% in the average group and decreased to 56.3% in the below-average and to 55.9% in the above-average group. Regarding the Criterion Nominal IQ, the comparability decreased to 44.4% in the below-average and to 46.1% in the above-average group. Regarding the Criterion Nominal CI95, the comparability decreased to 85.4% in the below-average and to 84.5% in the above-average group. For the latter two criteria, the comparability in the average group was 100% as it was artificially defined by group allocation.

Discussion

The main objective of the present study was to investigate comparability between IQs from seven widely used intelligence tests on the group level and in particular on the individual level. In line with Hypothesis 1, we found mainly substantial correlations between intelligence tests on the group level. Only 2 of 18 correlations were not significant, and these should be interpreted with caution, as sample sizes in these two cases were occasionally low. However, contrary to our Hypothesis 2 and to Hagmann-von Arx et al. (2018), we found that in 14 of 18 comparisons, the mean IQ differed significantly between two test scores. Investigating whether IQs obtained from different tests could be classified as comparable (exchangeable) on an individual level, we found that this was true for a low percentage that varied across the different criteria from 49.8% (criterion CI90) to 89.9% (criterion Nominal CI95). Within the percentage of incomparability, tests with earlier standardization years yielded higher IQs than tests with more recent standardization years, which we interpret as an indicator of the Flynn effect (Research Question 2). Beyond that, the nominal intelligence level (below- or above-average vs. average range) was the only systematic predictor of intelligence score differences between tests (Research Question 4). Therefore, we ran post hoc analyses on individual-level comparisons for the three nominal intelligence levels separately. With these analyses we not only were able to solve the issue regarding the bias toward a higher IQ in our study sample but also found that comparability was considerably lower in the

below- and above-average range (lowest comparability with 44.4% for criterion CI90) compared to the average range (lowest comparability with 61.4% for criterion CI90). This result is alarming and calls for explanation, as predominantly in these ranges diagnosis is particularly relevant for eligibility and high-stakes decisions. A similar pattern was found for the individual-level comparability of nonverbal and verbal factor scores as well (Research Question 3).

Individual-Level Comparisons

Before taking a closer look at the various criteria of comparability on the individual level, we would like to point out that we neither interpret nor compare comparability of specific test combinations (e.g., whether the comparability between the IDS-2 and the RIAS is higher than the comparability between the IDS-2 and the WISC-IV) because sample sizes varied considerably between test comparisons. Instead, we focus on the overall comparability in order to compare and discuss the advantages and disadvantages of the different criteria.

Our results show that comparability was lowest for the criterion CI90, as this interval was narrowest. With a between-test comparability of slightly more than 60% in the average and less than 50% in the below- and above-average IQ range, using the 90% CI for test interpretation does not seem to be useful. With the criterion proposed by Bergeron et al. (2008) to consider 5 IQ points above and below the exact point score when diagnosing intellectual ability (criterion IQ10), test comparability was higher in the average range with over 70% but still quite low in the below- and above-average range with about 50%. CI95 and nominal IQ have revealed the same percentage of comparability for the total sample with slightly over 60%. However, the CI95 led to better comparability in the below- and above-average range (around 55%) compared to the Nominal IQ (under 50%). Therefore, although the nominal IQ has practical relevance, as it allows for clear classification and may thus be used for diagnosis and classification (e.g., of intellectual disability or intellectual giftedness), comparability on this criterion is too low to be of diagnostic utility. The CI95 should hence be

preferred to the Nominal IQ. Last, the comparability between tests is maximized and approaches 90% when interpreted on the Nominal CI95 criterion. However, one must bear in mind that the IQ scores of two tests could differ by up to 45 points and still be classified as "comparable" according to this criterion. Given such a wide IQ range, hardly any useful decisions or conclusions can be made. For this reason, and even though this criterion may best represent the extent to which intelligence tests are incomparable, the Nominal CI95 does not provide a diagnostically useful interpretation of test results. Consequently, it seems very difficult to find a criterion that allows for both a reasonably narrow IQ range and a high probability that a comparable result would be achieved with another intelligence test. Overall, we conclude that for the IQ, the CI95 criterion has the best trade-off between comparability and accuracy. Still, the percentage of comparability, especially in the below- and above-average range, is disconcertingly low on this criterion too.

For the investigation of individual-level comparability of first-order factor scores, our results show that the comparability of NIs is lower compared to the comparability of IQs when reliability is *not* considered (criterion IQ10) and that the comparability of NIs does not differ from the comparability of IQs when reliability is considered (criterion CI95). However, the comparability of VIs was—although still quite low—significantly higher than that of IQs on both criteria, IQ10 and CI95. A possible explanation for the higher comparability of VIs could be that there is greater content overlap between the subtests of the different test batteries that constitute the VI compared to the NI and IQ. Although, content overlap should be neglectable for the measurement of *g* due to the principle of aggregation (Rushton et al., 1983), it does play an important role when the number of subtests is lower (Farmer et al., 2019), as it is in VI and NI. This could now lead to the conclusion that the VI is to be preferred to the IQ as an intelligence estimate. However, there is a large body of research with at least two widely supported counterarguments: First, IQ is a more accurate representation of *g* and first-order factor scores are usually less reliable than the IQ (Jensen, 1981, 1998). In this

sense, the finding that comparability is higher for VI than for IQ does not make VI a better representation of *g*. Second, factor scores contain only a little unique variance compared to the IQ (Canivez & Youngstrom, 2019; McGill et al., 2018). Consequently, we do not know to what extent this comparability is caused by shared variance of *g* and to what extent it is caused by more specific verbal reasoning. Therefore, we recommend interpreting NI as well as VI only with extreme caution, if at all, although it is a common practice among school psychologists. Individual-level comparability is only slightly higher for VI compared to IQ. With that, the discussed drawbacks of interpreting first-order factors cannot be overcome.

Moreover, the percentage of comparability is especially low in the below- and above-average range, not only for IQ, but also for NI and VI. At first glance it might seem plausible that the regression toward the mean explains this finding. However, this explanation does not withstand scrutiny, as the 95% CI was corrected for the regression toward the mean and comparability was still quite low. This not only calls for explanation but also shows the need to discuss what practitioners as well as test developers can do to deal with potential sources of incomparability.

Implications for the Application and Construction of Intelligence Tests

One way to deal with this comparability issue is to apply more than one intelligence test, especially for high-stakes decisions. In line with this, previous results from generalizability theory analyses showed that up to five intelligence tests need to be applied to achieve a reliability of at least .90 (Hagmann-von Arx et al., 2018). As this is obviously not feasible in practice, the authors suggested using at least two tests for high-stakes decisions. Yet, whereas the interpretation of a test result and subsequent diagnoses and decisions are strengthened if IQs obtained in two different tests are comparable, the question of what to do if they lie far apart remains unanswered. Whereas it seems to be common practice to average composite scores from different tests to gain a more reliable and valid estimation of true

intelligence, we do not support this recommendation as it may cause additional measurement error (Schneider & McGrew, 2011). Moreover, there are numerous reasons for possible underachievement (e.g., fatigue, sleep quality, motivational factors, test anxiety, environmental distractions, etc.) but rather few reasons for possible overachievement (e.g., guessing, learning effects, inappropriate help from the examiner) on intelligence tests. Measurement error can therefore not be averaged arithmetically. Instead of averaging IQs across tests if the scores diverge widely, we recommend considering anamnestic information such as language and cultural background and personal characteristic (e.g., interaction difficulties) as well as environmental conditions to determine if a child's and adolescent's abilities are accurately measured (American Psychological Association, 2017; International Test Commission, 2001; National Research Council, 2002). This might be a necessary step both prior to test administration for an adaptive test selection as well as after test administration for a careful test interpretation. Investigating some of these possible influences (i.e., bilingualism, age, and test interval) on test comparability, we could not detect any systematic effect. However, it is possible that additional factors that have not been examined in this study may have impacted test comparability on the individual level. Further, it is also possible that some potential factors influenced test comparability for certain individuals under certain circumstances, but not for others. If this was the case, we must assume that incomparability and measurement inaccuracy is increased by random variations (e.g., a temporary state of an individual or random circumstances) instead of systematic effects. This type of error variance is often called transient error (Schmidt, Le, & Ilies, 2003). The National Research Council (2002) provided a comprehensive list of possible systematic and transient error sources (threats) including characteristics of the examinee, examiner's influence, and environmental influences as well as psychometric test characteristics and stated that "all of them can be controlled to some considerable degree" (p. 100). Nevertheless, we note that this is a considerable challenge for practitioners, not only because of the multitude of potential

threats but also because of time pressure, that often plays a crucial role in the diagnostic process. Our results show the extent of incomparability across test scores when potential transient error sources are not considered. With this in mind, our results indicate how prone intelligence test scores are to interference, and how high the risk of misdiagnosis may be if the diagnostic process is not carried out with the utmost thoroughness. It thus seems important to explore what not only test administrators but also test developers can do to reduce potential sources of incomparability.

In this context, we consider it important to discuss the use of internal consistency coefficients for the calculation of CIs. This practice should be reconsidered for mainly two reasons: First, the specific internal consistency coefficient used for the calculation of CIs differs among tests (i.e., Cronbach's α or split-half reliability), which influences the width of the CI and with that also the comparability between tests. Second, internal consistency coefficients do not take into account the previously described transient errors. Yet, ignoring such transient errors leads to an overestimation of the reliability (Schmidt et al., 2003). Thus, it is important to investigate whether other reliability coefficients — such as the test–retest reliability that does consider transient errors — would provide an alternative and more accurate estimate and basis for the calculation of CIs. Nevertheless, this does not solve the problem of comparability seeming to be especially low in the below- and above-average intelligence levels, although CIs are already corrected for the regression toward the mean. If this finding is replicated in future research, it will be of interest to explore how and to what extent the IQ level should be considered within the calculation of reliability coefficients.

Generally, it is essential to ensure that comparability of intelligence test results is not reduced by sampling error, and that test developers proceed according to the current state of research. This includes making sure that the IQ is based on an adequate number of subtests (see Farmer et al., 2019), that the IQ consists of a heterogeneous set of subtests in order to average out measurement error, and that only subtests with a high *g* loading are included

(Farmer et al., 2019; Jensen & Weng, 1994). Further research is needed to investigate whether and how individual-level comparability might be enhanced with an optimal balance of these three aspects.

Limitations

Our study has several limitations. First, mean IQs of our sample were higher than the standardization means. This bias may have been caused by participants being allowed to decide whether they wanted to do more tests after having taken either the IDS-2 or the SB5. It is probable that participants with lower results on one of the first tests had less joy and were therefore less motivated to participate in further tests. However, and fortunately, this bias toward higher IQs could be addressed by splitting the sample into three nominal intelligence-level groups for individual-level analyses. Second, our sample sizes varied considerably between test comparisons. We therefore did not explicitly compare or interpret comparability of specific test combinations. Third, we had relatively small sample sizes of the youngest (4–6 years) and oldest (16–20 years) participants. This leads to a reduced variability and might therefore have impacted our regression analyses with age as a possible predictor of IQ differences between tests. Further research including more individuals in all age ranges might shed further light on the possible impact of age on test comparability. Finally, information about mono- or bilingualism was collected via self- or parent report. An examination of actual language skills by using performance-based tests could provide valuable information on whether and to what extent they influence test comparability.

Conclusion

We conclude that the interpretation of an exact IQ from a single intelligence test does not hold empirically, which strongly questions the use of cut-off values for diagnosis and related educational resource allocation. At least the 95% CI must be considered when making high-stakes decisions. Even then, and especially in the upper and lower bounds of the IQ scores, there is a considerable risk that diagnosis and resource allocation for special education

will be based largely on test-specific characteristics and measurement error and less on the individual's true ability. To reduce this risk and to enhance between-test comparability, diagnosticians are encouraged to consider possible factors that interfere with an individual's ability. Hence, test administrators need to be well trained and informed about the limitations of conventional intelligence tests. Further, our results point to the fact that current intelligence tests may have relied too much on results derived from group-level analyses while excluding individual-level analyses to determine psychometric test quality. Future research is needed to generate effective approaches to enhancing accuracy and with this also individual-level comparability between tests.

References

Alexander, R. A., Carson, K. P., Alliger, G. M., & Carr, L. (1987). Correcting doubly truncated correlations: An improved approximation for correcting the bivariate normal correlation when truncation has occurred on both variables. *Educational and Psychological Measurement*, *47*, 309–315. https://doi.org/10.1177/0013164487472002

Allen, D. N., Stolberg, P. C., Thaler, N. S., Sutton, G., & Mayfield, J. (2014). Validity of the RIAS for assessing children with traumatic brain injury: Sensitivity to TBI and comparability to the WISC-III and WISC-IV. *Applied Neuropsychology: Child*, *3*, 83–93. https://doi.org/10.1080/21622965.2012.700531

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct, including 2010 and 2016 amendments*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Aspara, J., Wittkowski, K., & Luo, X. (2018). Types of intelligence predict likelihood to get married and stay married: Large-scale empirical evidence for evolutionary theory. *Personality and Individual Differences*, *122*, 1–6. https://doi.org/10.1016/j.paid.2017.09.028

Atkinson, L. (1989). Three standard errors of measurement and the Stanford-Binet Intelligence Scale, Fourth Edition. *Psychological Assessment*, *1*, 242–244. https://doi.org/http://dx.doi.org/10.1037/1040-3590.1.3.242

Baum, K. T., Shear, P. K., Howe, S. R., & Bishop, S. L. (2015). A comparison of WISC-IV and SB-5 intelligence scores in adolescents with autism spectrum disorder. *Autism*, *19*, 736–745. https://doi.org/10.1177/1362361314554920

Benson, N., Floyd, R. G., Kranzler, J. H., Eckert, T. L., & Fefer, S. (2018). Contemporary assessment practices in school psychology: National survey results. *Paper Presented at*

the Meeting of the National Association of School Psychologists, Chicago, IL.

Bergeron, R., Floyd, R. G., & Shands, E. I. (2008). States' eligibility guidelines for mental retardation: An update and consideration of part scores and unreliability of IQs. *Education and Training in Developmental Disabilities*, *43*, 123–131.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*, 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Calvin, C. M., Deary, I. J., Fenton, C., Roberts, B. A., Der, G., Leckenby, N., & Batty, G. D. (2011). Intelligence in youth and all-cause-mortality: Systematic review with meta-analysis. *International Journal of Epidemiology*, *40*, 626–644. https://doi.org/10.1093/ije/dyq190

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical, clinical, and policy implications. *Applied Measurement in Education*, *32*, 232–248. https://doi.org/10.1080/08957347.2019.1619562

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Damian, R. I., Su, R., Shanahan, M., Trautwein, U., & Roberts, B. W. (2015). Can personality traits and intelligence compensate for background disadvantage? Predicting status attainment in adulthood. *Journal of Personality and Social Psychology*, *109*, 473–489. https://doi.org/10.1037/pspp0000024.Can

Daseking, M., Lipsius, M., Petermann, F., & Waldmann, H. C. (2008). Differenzen im Intelligenzprofil bei Kindern mit Migrationshintergrund: Befunde zum HAWIK-IV [Intelligence and cultural influences: Differences in the intelligence profile of children with a migration background: Findings on WISC-IV]. *Kindheit und Entwicklung*, *17*, 76–89. https://doi.org/10.1026/0942-5403.17.2.76

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational

achievement. *Intelligence*, *35*, 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*, 7716–7720. https://doi.org/10.1073/pnas.1018601108

Dutton, E., & Lynn, R. (2014). A negative Flynn effect in Finland, 1997-2009. *Intelligence*, *41*(6), 817–820. https://doi.org/10.1016/j.intell.2013.05.008

Evers, A. (2001). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, *1*, 137–153. https://doi.org/10.1207/S15327574IJT0102

Farmer, R. L., Floyd, R. G., Reynolds, M. R., & Berlin, K. S. (2019). How can general intelligence composites most accurately index psychometric *g* and what might be good enough ? *Contemporary School Psychology*. https://doi.org/10.1007/s40688-019-00244-1

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice*, *39*, 414–423. https://doi.org/10.1037/0735-7028.39.4.414

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 427–427. https://doi.org/10.1037/h0090408

Flynn, J. R. (2009). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938-2008. *Economics and Human Biology*, *7*, 18–27. https://doi.org/10.1016/j.ehb.2009.01.009

Gallinat, E., & Spaulding, T. J. (2014). Differences in the performance of children with specific language impairment and their typically developing peers on nonverbal cognitive tests: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *57*, 1363–1382. https://doi.org/10.1044/2014

Goldstein, D., Hahn, C. S., Hasher, L., Wiprzycka, U. J., & Zelazo, P. D. (2007). NIH Public Access. *Personality and Individual Differences*, *42*, 431–440.

https://doi.org/doi:10.1016/j.paid.2006.07.008

Grob, A., Gygi, J. T., & Hagmann-von Arx, P. (2019). *The Stanford-Binet Intelligence Scales, Fifth Edition (SB5)—German Adaptation*. Bern, Switzerland: Hogrefe.

Grob, A., & Hagmann-von Arx, P. (2018). *Intelligence and Development Scales–2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche [Intelligence and Development Scales for children and adolescents]*. Bern, Switzerland: Hogrefe.

Grob, A., Reiman, G., Gut, J., & Frischknecht, M. C. (2013). *Intelligence and Development Scales – Preschool (IDS-P). Intelligenz- und Entwicklungsskalen für das Vorschulalter [Intelligence and Development Scales for Children from 5-10 years of age]*. Bern: Hans Huber.

Gupta, S. (1991). Effects of time of day and personality on intelligence test scores. *Personality and Individual Differences*, *12*, 1227–1231. https://doi.org/10.1016/0191-8869(91)90089-T

Hagmann-von Arx, P., & Grob, A. (2014). *Reynolds Intellectual Assessment Scales (RIAS)—German adaptation*. Bern, Switzerland: Hans Huber.

Hagmann-von Arx, P., Grob, A., Petermann, F., & Daseking, M. (2012). Konkurrente Validität des HAWIK-IV und der Intelligence and Development Scales (IDS) [Concurrent validity of the HAWIK-IV and the Intelligence and Development Scales (IDS)]. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *40*, 41–50. https://doi.org/10.1024/1422-4917/a000148

Hagmann-von Arx, P., Lemola, S., & Grob, A. (2018). Does IQ = IQ ? Comparability of intelligence test scores in typically developing children. *Assessment*, *25*, 691–701. https://doi.org/10.1177/1073191116662911

Hagmann-von Arx, P., Petermann, F., & Grob, A. (2013). Konvergente und diskriminante Validität der WISC-IV und der Intelligence and Development Scales (IDS) bei Kindern mit Migrationshintergrund [Convergent and discriminant validity of the WISC-IV and

the Intelligence and Development Scales (IDS) in children . *Diagnostica*, *59*, 170–182. https://doi.org/10.1026/0012-1924/a000091

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*, 93–114.

Jensen, A. R. (1981). *Straight talk about mental tests*. New York, NY: Free Press.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jensen, A. R., & Weng, L.-J. (1994). What is a good g? *Intelligence*, *258*, 231–258.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General Intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology*, *86*, 96–111. https://doi.org/10.1037/0022-3514.86.1.96

Lynn, R., & Harvey, J. (2008). The decline of the world's IQ. *Intelligence*, *36*, 112–120. https://doi.org/10.1016/j.intell.2007.03.004

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. https://doi.org/10.1016/j.jsp.2018.10.007

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–181). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–182). New York, NY: Guilford

Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1–10.

Miller, C. A., & Gilbert, E. (2008). Comparison of performance on two nonverbal intelligence tests by adolescents with and without language impairment. *Journal of Communication Disorders*, *41*, 358–371. https://doi.org/10.1016/j.jcomdis.2008.02.003

National Research Council. (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academy Press.

Neukrug, E., & Fawcett, C. (2014). *Essentials of testing and assessment: A practical guide for counselors, social workers, and psychologists.* Belmont, CA: Thomson Brooks/Cole.

Petermann, F. (2009). *Wechsler Preschool and Primary Scale of Intelligence–Third Edition (WPPSI-III; German Adaptation)*. Frankfurt am Main, Germany: Pearson Assessment.

Petermann, F., & Petermann, U. (2011). *Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; German adaptation)*. Frankfurt am Main, Germany: Pearson Assessment.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*, 18–38. https://doi.org/10.1037/0033-2909.94.1.18

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206–224. https://doi.org/10.1037/1082-989X.8.2.206

Schneider, W. J., & McGrew, K. S. (2011). *Just say no" to averaging IQ subtest scores* (applied psychometrics 101, Technical Report #10). https://doi.org/10.13140/RG.2.2.19863.37289

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A

study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence*, *32*, 349–362. https://doi.org/10.1016/j.intell.2004.06.004

Tellegen, P. J., Laros, J. A., & Petermann, F. (2012). *SON-R 6-40. Snijders-Oomen Nonverbaler Intelligenztest [SON-R 6-40 Snijders-Oomen Nonverbal Intelligence Test]*. Göttingen, Germany: Hogrefe.

von Aster, M. G., Neubauer, A., & Horn, R. (2006). *Wechsler Intelligenztest für Erwachsene (WIE-III) [German adaptation of the Wechsler Adult Intelligence Scale III]*. Frankfurt am Main, Germany: Harcourt Test Services.

World Health Organization [WHO]. (2018). International Classification of Diseases for Mortality and Morbidity Statistics (11th rev.). Retrieved from https://icd.who.int/browse11/l-m/en

Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, *33*, 292–296. https://doi.org/10.1037/a0030727

Yin Foo, R., Guppy, M., & Johnston, L. M. (2013). Intelligence assessments for children with cerebral palsy: A systematic review. *Developmental Medicine and Child Neurology*, *55*, 911–918. https://doi.org/10.1111/dmcn.12157

Table 1

Test Characteristics

Test	Age range (Years;months)	Factors and indices	Standardization years	Internal consistency[a]		
				FSIQ	NI	VI
IDS-2	5;0 to 20;11	Visual processing, processing speed, auditory short-term memory, visuospatial short-term memory, long-term memory, abstract reasoning[b], verbal reasoning[c], general intelligence[d]	2015–2017	.98	.94	.96
SB5	4;0 to 80;11	Fluid reasoning, knowledge, quantitative reasoning, visuospatial processing, working memory, verbal intelligence[c], nonverbal intelligence[b], general intelligence[d]	2015–2017	.99	.98	.99
RIAS	3;0 to 99;11	Verbal intelligence[c], nonverbal intelligence[b], general intelligence[d]	2011–2012	.95	.93	.94
SON-R 6-40	6;0 to 40;11	Nonverbal intelligence[b,d]	2009–2011	.95	.95	N/A
WAIS-III	16;0 to 89;11	Verbal intelligence[c] (verbal comprehension, working memory), performance intelligence[b] (perceptual organization, processing speed), general intelligence[d]	1999–2005	.97	.94	.96
WISC-IV	6;0 to 16;11	Verbal comprehension[c], perceptual reasoning[b], working memory, processing speed, general intelligence[d]	2005–2006	.97	.93	.94
WPPSI-III	3;0 to 7;2	Verbal intelligence[c], performance intelligence[b], processing speed, general language, general intelligence[d]	2001–2002	.95	.91	.92

Note. FSIQ = Full-scale IQ; NI = Nonverbal Intelligence Index; VI = Verbal Intelligence Index; IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; SON-R 6-40 = Snijders–Oomen Nonverbal Intelligence Test 6-40, German adaptation; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition, German adaptation; WISC-IV = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation; WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation.

[a] Reliability is reported as internal consistency coefficients. According to the corresponding test manuals, coefficients of the IDS-2, RIAS, and WISC-IV are Cronbach's α and coefficients of the SB5, WAIS-III, and WPPSI-III are split-half reliabilities.

[b] Index used for nonverbal intelligence analyses.

[c] Index used for verbal intelligence analyses.

[d] Index used for FSIQ analyses.

Table 2

Descriptive Statistics and One-Sample t *Tests*

Test	*N*	Minimum	Maximum	*M*	*SD*	Skewness	Kurtosis	*t*/*Z*[a]
IDS-2	231	69	133	102.46	12.05	-0.129	-0.001	3.11**
SB5	202	55	134	102.47	12.80	-0.397	0.721	2.74**
RIAS	327	76	132	105.03	9.64	-0.057	0.305	9.44***
SON-R 6-40	269	60	145	113.91	15.17	-0.120	-0.023	15.03***
WAIS-III	34	77	136	115.29	12.57	-0.685	1.351	7.09***
WISC-IV	226	60	142	112.49	11.59	-0.711	2.255	11.37***
WPPSI-III	46	78	134	105.70	12.13	0.028	-0.141	3.19**

Note. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; SON-R 6-40 = Snijders–Oomen Nonverbal Intelligence Test 6-40, German adaptation; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition, German adaptation; WISC-IV = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation; WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation.

[a] Comparisons including the WISC-IV were analyzed with the Wilcoxon sign-ranked test (*Z*).

** $p < .01$; *** $p < .001$.

Table 3

Group-Level Comparison of IQs

Test comparison	Difference in standardization years	*N*	r/ρ[a]	r_{VarRel}	Mean differences (*SD*)	*t*/*Z*[b]	*d*/*r*[c]
IDS-2 and SB5	0	55	.73***	.82***	0.62 (9.90)	0.46	0.05
IDS-2 and RIAS	5	196	.65***	.81***	-3.04 (9.58)	-4.44***	-0.27
IDS-2 and SON-R 6-40	6	154	.67***	.74***	-10.44 (11.02)	-11.75***	-0.77
IDS-2 and WAIS-III	12	30	.46*	.59***	-15.33 (12.39)	-6.78***	-1.29
IDS-2 and WISC-IV	11	113	.65***	.78***	-8.66 (9.31)	-7.52***	-0.71
IDS-2 and WPPSI-III	15	24	.87***	.94***	-4.54 (5.96)	-3.74**	-0.39
SB5 and RIAS	5	167	.60***	.76***	-2.00 (10.42)	-2.48*	-0.17
SB5 and SON-R 6-40	6	139	.71***	.76***	-12.63 (11.09)	-13.43***	-0.87
SB5 and WAIS-III	12	11	.28	.38	-8.55 (12.91)	-2.20	-0.80
SB5 and WISC-IV	11	138	.75***	.84***	-10.32 (7.90)	-9.49***	-0.81
SB5 and WPPSI-III	15	29	.68***	.78***	-4.34 (10.10)	-2.32*	-0.35
RIAS and SON-R 6-40	1	248	.48***	.63***	-8.59 (13.16)	-10.28***	-0.67
RIAS and WAIS-III	7	28	.56**	.74***	-10.43 (11.24)	-4.91***	-0.87
RIAS and WISC-IV	6	199	.66***	.83***	-7.34 (8.59)	-9.32***	-0.66
RIAS and WPPSI-III	10	30	.51**	.71***	1.23 (11.76)	0.57	0.10
SON-R 6-40 and WAIS-III	6	24	.31	.37	-7.00 (15.57)	-2.20*	-0.53
SON-R 6-40 and WISC-IV	5	171	.61***	.70***	2.81 (10.86)	-3.05**	-0.23
SON-R 6-40 and WPPSI-III	9	18	.41	.49*	9.28 (13.79)	2.85*	0.73

Note. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; SON-R 6-40 = Snijders–Oomen Nonverbal Intelligence Test 6-40, German adaptation; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition, German adaptation; WISC-IV = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation; WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation. r_{VarRel} = Correlations corrected for range restriction and attenuation.

[a] Comparisons including the WISC-IV were analyzed using the nonparametric Spearman rank-order correlation (ρ). All other comparisons were analyzed using the Pearson product-moment correlation (r).

[b] Comparisons including the WISC-IV were analyzed with the Wilcoxon test (Z). All other comparisons were analyzed with the dependent t test (t).

[c] Effect sizes including the WISC-IV were indicated as $r = Z/\sqrt{N}$. All other effect sizes are indicated as Cohen's d.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4

Individual-Level Comparison of IQs

Test comparison	*N*	M_{diff}	Critical difference[a]	CI90[b] (%)	Higher intelligence[c] (%)	Lower intelligence[d] (%)	IQ10[e] (%)	CI95[f] (%)	Nominal IQ[g] (%)	Nominal CI95[h] (%)
IDS-2 and SB5	55	7.31	5.94	52.7	30.9	16.4	78.2	61.8	69.1	89.1
IDS-2 and RIAS	196	7.99	8.98	63.8	9.2	27.0	72.4	74.5	76.0	96.9
IDS-2 and SON-R 6-40	154	12.36	8.98	40.3	4.5	55.2	48.7	53.2	55.8	85.1
IDS-2 and WAIS-III	30	15.60	8.40	30.0	0.0	70.0	33.3	33.3	53.3	90.0
IDS-2 and WISC-IV	113	10.26	7.74	43.4	3.5	53.1	59.3	56.6	63.7	92.0
IDS-2 and WPPSI-III	24	6.13	8.98	79.2	4.2	16.6	87.5	91.7	75.0	100.0
SB5 and RIAS	167	8.17	7.96	55.7	16.8	27.5	70.7	67.7	77.2	92.2
SB5 and SON-R 6-40	139	13.78	7.96	33.1	2.9	64.0	41.7	41.7	53.2	79.9
SB5 and WAIS-III	11	11.09	7.38	36.4	9.1	54.5	63.6	45.5	63.6	81.8
SB5 and WISC-IV	138	11.07	6.72	29.0	2.2	68.8	53.6	43.5	55.8	89.9
SB5 and WPPSI-III	29	9.59	7.96	31.0	20.7	48.3	58.6	62.1	65.5	96.6
RIAS and SON-R 6-40	248	12.38	11.00	54.0	7.3	38.7	50.0	64.1	56.5	85.1
RIAS and WAIS-III	28	12.79	10.42	39.3	7.1	53.6	39.3	67.9	46.4	89.3
RIAS and WISC-IV	199	9.44	9.76	55.3	4.0	40.7	59.8	68.8	57.8	93.5
RIAS and WPPSI-III	30	9.03	11.00	76.7	16.7	6.6	70.0	80.0	66.7	93.3
SON-R 6-40 and WAIS-III	24	12.67	10.42	58.3	12.5	29.2	58.3	62.5	50.0	83.3
SON-R 6-40 and WISC-IV	171	9.10	9.76	57.3	28.7	14.0	60.2	71.3	60.8	92.4
SON-R 6-40 and WPPSI-III	18	13.61	11.00	44.4	44.4	11.2	38.9	61.1	55.6	88.9
Total sample	1774	10.45	8.96	49.8[i]	9.6[i]	39.9	58.1	61.9	61.9	89.9
Total below average	144	11.77	8.85	44.4[i]	9.5[i]	46.1	50.7	56.3	44.4	85.4
Total average	608	7.62	8.98	61.4[i]	8.1[i]	30.5	71.2	75.6	100	100
Total above average	972	12.10	9.12	44.4[i]	11.2[i]	44.4	51.5	55.9	41.6	84.5

Note. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; SON-R 6-40 = Snijders–Oomen Nonverbal Intelligence Test 6-40, German adaptation; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition, German adaptation; WISC-IV = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation; WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation. CI = confidence interval; Total above sample = participants who scored in the above average or upper extreme range on at least one intelligence test; Total average = participants who scored in the average range on all intelligence tests; Total below average = participants who scored in the below average or lower extreme range on at least one intelligence test.

[a] The sum of half of each test's 90% confidence interval in intelligence test scores.

[b] The percentage of participants who reached a difference between each pair of intelligence test scores of less than or equal to the critical difference.

[c] The percentage of participants who scored higher on the first-listed and more recently standardized intelligence test of each pair of intelligence tests.

[d] The percentage of participants who scored lower on the first-listed and more recently standardized intelligence test of each pair of intelligence tests.

[e] The percentage of participants who reached a difference between each pair of intelligence test scores of less than or equal to 10 IQ points.

[f] The percentage of participants with overlapping 95% CIs.

[g] The percentage of participants who scored on the same qualitative nominal intelligence level in both tests (<70: lower extreme; 70–84: below average; 85–115: average; 116–130: above average; >130: upper extreme).

[h] The percentage of participants who scored on the same qualitative nominal intelligence level in both tests when considering both qualitative nominal intelligence levels if the 95% CI spanned two levels (e.g., above average to average).

[i] This mean percentage was calculated without the comparison of the IDS-2 and SB5, as they were standardized within the same period.

Table 5

Linear Regressions With Age, Bilingualism, Time Interval Between Tests, Order of Test Administration, and Nominal Intelligence Level as Predictors of IQ Differences Between Two Intelligence Tests

Test comparison	IDS-2 with						SB5 with					RIAS with				SON-R 6-40 with		
	SB5	RIAS	SON-R 6-40	WAIS-III	WISC-IV	WPPSI-III	RIAS	SON-R 6-40	WAIS-III	WISC-IV	WPPSI-III	SON-R 6-40	WAIS-III	WISC-IV	WPPSI-III	WAIS-III	WISC-IV	WPPSI-III
Age	.20	.01	-.08	.25	.14	-.31	-.05	-.16	.32	.26**	-.55**	-.20**	.07	-.10	-.35	.08	-.16*	-.11
Bilingualism	-.19	-.06	.02	-.29	.01	.18	.16*	.16	-.34	-.07	.45*	.03	-.92	-.04	.04	-.11	.18*	.37
Time interval between tests	-.08	.07	.07	.20	.19*	.25	.13	.14	.18	-.08	.14	-.03	-.02	.13	.09	.05	.05	-.14
Order of test administration	.09	.09	.17*	.12	.19*	-.14	.11	.03	.37	-.03	.15	.02	.32	-.02	.13	.08	.05	-.30
Nominal intelligence level																		
Below average	.45**	.29***	.14	.18	.14	-.31	.39***	.26**	NA	.37***	.57**	.02	.39	.06	.06	.27	-.01	NA
Above average	.24	.18**	.50***	.70***	.25**	-.59**	.31***	.54***	.51	.22**	.46**	.56***	.53**	.41***	.45*	.62**	.32***	.55*

Note. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; SON-R 6-40 = Snijders–Oomen Nonverbal Intelligence Test 6-40, German adaptation; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition, German adaptation; WISC-IV = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation; WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation. Coefficients are standardized beta regression coefficients. Bilingualism was coded as follows: 1 = female; 0 = male. Order of test administration was coded as follows: 1 = the first listed test was conducted earlier; 0 = the first listed test was conducted later. Nominal intelligence level: below average = IQ < 85; above average = IQ > 115; average (IQ 85–115) was the reference category. NA = analyses could not be completed as none of the participants that completed either of the test combinations achieved an IQ in the below-average range.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Supplemental Table 1

Individual-Level Comparison of Nonverbal Factor Scores

Test comparison	*N*	M_{diff}	SD_{diff}	$Diff_{min}$	$Diff_{max}$	IQ10[a] (%)	CI95[b] (%)
$IDS\text{-}2_{AR}$ and $SB5_{NI}$	55	11.9	9.5	0	38	52.7	58.2
$IDS\text{-}2_{AR}$ and $RIAS_{NI}$	196	11.0	8.1	0	46	53.1	72.4
$IDS\text{-}2_{AR}$ and SON-R 6-40	154	13.0	10.3	0	46	50.0	59.7
$IDS\text{-}2_{AR}$ and $WAIS\text{-}III_{PI}$	30	19.7	12.1	2	44	26.7	33.3
$IDS\text{-}2_{AR}$ and $WISC\text{-}IV_{PR}$	114	14.3	10.2	0	46	43.9	58.8
$IDS\text{-}2_{AR}$ and $WPPSI\text{-}III_{PI}$	24	14.3	9.4	1	34	41.7	41.7
$SB5_{NI}$ and $RIAS_{NI}$	168	11.4	9.1	0	51	56.5	63.1
$SB5_{NI}$ and SON-R 6-40	140	14.6	9.8	0	47	37.9	40.0
$SB5_{NI}$ and $WAIS\text{-}III_{PI}$	11	12.5	14.6	1	46	63.6	63.6
$SB5_{NI}$ and $WISC\text{-}IV_{PR}$	139	15.4	10.1	0	43	36.7	42.4
$SB5_{NI}$ and $WPPSI\text{-}III_{PI}$	29	11.5	10.1	0	39	55.2	58.6
$RIAS_{NI}$ and SON-R 6-40	248	13.1	10.0	0	53	49.6	64.1
$RIAS_{NI}$ and $WAIS\text{-}III_{PI}$	28	15.4	10.9	0	47	39.3	60.7
$RIAS_{NI}$ and $WISC\text{-}IV_{PR}$	200	12.2	8.8	0	42	48.0	71.5
$RIAS_{NI}$ and $WPPSI\text{-}III_{PI}$	30	11.6	8.9	1	33	60.0	66.7
SON-R 6-40 and $WAIS\text{-}III_{PI}$	24	11.9	12.0	1	43	66.7	66.7
SON-R 6-40 and $WISC\text{-}IV_{PR}$	172	9.0	6.3	0	33	61.0	83.1
SON-R 6-40 and $WPPSI\text{-}III_{PI}$	18	11.9	8.5	1	28	44.4	66.7
Total sample	1,780	12.61	9.29	0	53	49.3	62.2
Total below average	144	14.56	9.51	0	44	37.5	50.7
Total average	611	10.00	7.22	0	36	59.9	72.0
Total above average	1,037	13.93	9.69	0	53	44.4	57.7

Note. $IDS\text{-}2_{AR}$ = Intelligence and Development Scales–2 Abstract Reasoning; $SB5_{NI}$ = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation, nonverbal intelligence; $RIAS_{NI}$ = Reynolds Intellectual Assessment Scales, German adaptation, nonverbal intelligence; SON-R 6-40 = Snijders–Oomen Nonverbal Intelligence Test 6-40, German adaptation; $WAIS\text{-}III_{PI}$ = Wechsler Adult Intelligence Scale, German adaptation, performance intelligence; $WISC\text{-}IV_{PR}$ = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation, perceptual reasoning; $WPPSI\text{-}III_{PI}$ = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation, performance intelligence. CI = confidence interval; total above sample = participants who scored in the above-average or upper extreme range on at least one intelligence test; total average = participants who scored in the average range on all intelligence tests; total below average = participants who scored in the below-average or lower extreme range on at least one intelligence test.

[a] The percentage of sample participants who reached a difference between each pair of intelligence test scores of less than or equal to 10 IQ points.

[b] The percentage of sample participants with overlapping 95% confidence intervals.

Supplemental Table 2

Individual-Level Comparison of Verbal Factor Scores

Test comparison	*N*	M_{diff}	SD_{diff}	$Diff_{min}$	$Diff_{max}$	IQ10[a] (%)	CI95[b] (%)
$IDS\text{-}2_{VR}$ and $SB5_{VI}$	55	8.42	7.09	0	32	70.9	61.8
$IDS\text{-}2_{VR}$ and $RIAS_{VI}$	196	8.41	6.63	0	30	69.4	79.1
$IDS\text{-}2_{VR}$ and $WAIS\text{-}III_{VI}$	30	13.17	7.41	0	29	43.3	53.3
$IDS\text{-}2_{VR}$ and $WISC\text{-}IV_{VC}$	113	10.60	8.41	0	45	56.6	64.6
$IDS\text{-}2_{VR}$ and $WPPSI\text{-}III_{VI}$	24	7.50	7.33	1	27	83.3	83.3
$SB5_{VI}$ and $RIAS_{VI}$	167	8.72	6.21	0	27	66.5	66.5
$SB5_{VI}$ and $WAIS\text{-}III_{VI}$	11	11.64	7.24	0	25	45.5	45.5
$SB5_{VI}$ and $WISC\text{-}IV_{VC}$	138	11.37	8.25	0	42	53.6	53.6
$SB5_{VI}$ and $WPPSI\text{-}III_{VI}$	29	9.07	6.09	0	22	58.6	62.1
$RIAS_{VI}$ and $WAIS\text{-}III_{VI}$	28	9.86	6.46	1	25	60.7	71.4
$RIAS_{VI}$ and $WISC\text{-}IV_{VC}$	199	9.82	8.12	0	46	62.3	73.9
$RIAS_{VI}$ and $WPPSI\text{-}III_{VI}$	30	8.10	5.67	0	24	70.0	90.0
Total sample	1,020	9.58	7.29	0	46	62.8	68.6
Total below average	90	10.79	8.02	0	39	57.8	61.1
Total average	358	8.28	6.21	0	34	71.8	77.1
Total above average	578	10.21	7.61	0	46	58.1	64.5

Note. $IDS\text{-}2_{VR}$ = Intelligence and Development Scales–2 verbal reasoning; $SB5_{VI}$ = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation, verbal intelligence; $RIAS_{VI}$ = Reynolds Intellectual Assessment Scales, German adaptation, verbal intelligence; $WAIS\text{-}III_{VI}$ = Wechsler Adult Intelligence Scale, German adaptation, verbal intelligence; $WISC\text{-}IV_{VC}$ = Wechsler Intelligence Scales for Children–Fourth Edition, German adaptation, verbal comprehension; $WPPSI\text{-}III_{VI}$ = Wechsler Preschool and Primary Scale of Intelligence–Third Edition, German adaptation, verbal intelligence. CI = confidence interval; Total above sample = participants who scored in the above-average or upper extreme range on at least one intelligence test; Total average = participants who scored in the average range on all intelligence tests; Total below average = participants who scored in the below-average or lower extreme range on at least one intelligence test.

[a] The percentage of sample participants who reached a difference between each pair of intelligence test scores of less than or equal to 10 IQ points.

[b] The percentage of sample participants with overlapping 95% confidence intervals.

APPENDIX B: Study II

Grieder, S., Bünger, A., Odermatt, S. D., Schweizer, F., & Grob, A. (2019). Limited individual-level comparability of IQs within three test batteries: Impact on external validity, some explanations, and possible solutions. *Manuscript submitted to the Journal of Educational Psychology*.

Draft January 20, 2020

Limited Individual-Level Comparability of IQs Within Three Test Batteries: Impact on External Validity, Some Explanations, and Possible Solutions

Silvia Grieder, Anette Bünger, Salome D. Odermatt, Florine Schweizer, and Alexander Grob

University of Basel

Author Note

Silvia Grieder, Anette Bünger, Salome D. Odermatt, Florine Schweizer, Alexander Grob, Department of Psychology, University of Basel, Switzerland.

Correspondence concerning this article should be addressed to Silvia Grieder, Department of Psychology, University of Basel, Missionsstrasse 62, 4055 Basel, Switzerland. E-mail: silvia.grieder@unibas.ch

Author contributions: Silvia Grieder provided conceptualization, methodology, formal analysis, writing–original draft, writing–review and editing, and visualization; Anette Bünger provided conceptualization and writing–review and editing; Salome D. Odermatt provided writing–review and editing; Florine Schweizer provided writing–review and editing; Alexander Grob provided conceptualization, resources, writing–review and editing, and supervision.

Acknowledgments: We thank Anita Todd for copy editing.

Abstract

Research on comparability of general intelligence composites (IQs) is scarce and has focused exclusively on comparing IQs from different test batteries, revealing limited individual-level comparability. We add to these findings, having, for the first time, investigated the group- and individual-level comparability of different IQs from the same test battery. Thereby, transient error was held to a minimum and between-battery variance was ruled out completely. We (a) determined the magnitude of intraindividual IQ differences, (b) investigated their impact on external validity (i.e., relationships with school grades), (c) explored possible predictors for these differences, and (d) examined possible solutions to incomparability in the form of confidence intervals (CIs) based on different reliability coefficients. Results are based on the standardization samples of three intelligence test batteries, spanning from early childhood to late adulthood. Despite high group-level comparability, individual-level comparability was often unsatisfactory, especially towards the tails of the IQ distribution. This limited comparability has consequences for external validity, as IQs were differentially related to and often less predictive for school grades for individuals with high intraindividual IQ differences. Of several predictors, only IQ level and age were systematically related to comparability, suggesting these attributes should be considered in the calculation of CIs. Consequently, findings challenge the use of an overall internal consistency for CIs and suggest using CIs based on test–retest reliability or age- and IQ-specific internal consistencies for clinical interpretation. Implications for the construction and application of intelligence tests are discussed.

Keywords: general intelligence, IQ, screening, individual level, reliability, validity

Limited Individual-Level Comparability of IQs Within Three Test Batteries: Impact on External Validity, Some Explanations, and Possible Solutions

General intelligence is defined as the broad mental capacity to reason, solve problems, comprehend complex ideas, and learn quickly (Gottfredson, 1997). It predicts numerous important life outcomes, including academic achievement (Lubinski, 2004; Roth et al., 2015), occupational success, socioeconomic status, income (Batty, Gale, Tynelius, Deary, & Rasmussen, 2009; Gottfredson, 2004; Lubinski, 2004), health, and longevity (Batty et al., 2009; Gottfredson & Deary, 2004).

The concept of general intelligence was first introduced by Charles Spearman as a common factor explaining the positive manifold of cognitive test outcomes—psychometric *g* (Spearman, 1904). Since Spearman, research on intelligence structure has moved to hierarchical models, but the majority of these models still includes a superordinate *g* factor. The currently perhaps most influential intelligence model, the Cattell–Horn–Carroll (CHC) model (Alfonso, Flanagan, & Radwan, 2005; McGrew, 1997, 2009), assumes a three-stratum structure with narrow abilities (or subtests) at the bottom that are indicators of broad abilities (or group factors, such as fluid reasoning, comprehension knowledge, perceptual speed), which are in turn influenced by *g*. Although the existence of a *g* factor is open to debate in the CHC taxonomy (e.g., McGrew, 2009), virtually all intelligence tests whose development was based on the CHC model—and almost all intelligence tests in general—include a Full-Scale IQ (FSIQ) as an indicator of general intelligence. To avoid an intertwining of the theoretical construct of general intelligence and its measurement, we refer to the theoretical construct as *general intelligence*, to the latent measure of general intelligence as *g* or the *g factor*, and to a (unit-weighted) subtest composite intended to measure general intelligence as *IQ*.

As most intelligence tests include a general intelligence measure, a major question in test construction concerns the determinants of a reliable and valid measurement of general intelligence. A recent study by Farmer, Floyd, Berlin, and Reynolds (2019) investigated such determinants by comparing the reliability and accuracy of different

intelligence composites from two test batteries. A composite's accuracy was defined as the magnitude of its loading on the *g* factor of the other test battery. Farmer et al. systematically varied the heterogeneity, *g* loadings (both separately and in combination), and number of subtests and found that, as a single criterion, high *g* loadings were more important than heterogeneity for an accurate composite. The most accurate composites were those derived from numerous (12 to 13) highly *g*-loaded and diverse subtests. However, their results also suggest a diminishing marginal utility concerning the number of subtests, as the gains in reliability and accuracy begin to flatten out from about four subtests on. Yet, as the authors pointed out, small gains in reliability are of practical relevance, as they can have substantial effects on confidence intervals (CIs) and hence on comparability on an individual level (i.e., overlap of CIs). It is therefore important to investigate individual-level in addition to group-level comparability to gain further insights into the accuracy of different composites.

Such individual-level comparisons have been performed between IQs derived from different test batteries (Bünger, Grieder, Schweizer, & Grob, 2019; Floyd, Clark, & Shadish, 2008; Hagmann-von Arx, Lemola, & Grob, 2018). These revealed substantial intraindividual absolute differences in IQs—henceforth called IQ differences—and limited comparability of CIs and IQs in nominal categories. All three of the aforementioned studies concluded that any two intelligence tests do not necessarily render comparable FSIQs on the individual level, even if they show high correlations and no mean differences on the group level.

We add to these previous findings with the present study, in which we investigated the individual-level comparability of different IQs derived from the same test battery that are all intended to measure general intelligence. Proceeding this way, transient error (i.e., errors due to variations in mood, information-processing efficiency, etc. over time; see Schmidt, Le, & Ilies, 2003) is kept to a minimum, and between-battery variance (i.e., differences in global test characteristic, such as general instructions, type of presentation, general appearance of test material) is held constant. For this purpose, the comparison between the FSIQ and a Screening IQ (ScrIQ) from the same test

battery is well suited. While the FSIQ is typically based on a larger number of subtests or all subtests from a test battery, a ScrIQ is based on a subset of subtests and is intended as a rough estimate of an individual's intellectual potential if time constraints are high.

After practitioners have administered a screening, their decision as to whether the rest of the test battery will also be administered is often based on the screening results (Thompson, LoBello, Atkinson, Chisholm, & Ryan, 2004). It is therefore especially important to investigate the individual-level comparability of the FSIQ and the ScrIQ. To this end, we first determined the magnitude of intraindividual differences in IQs; second, we investigated the impact of these differences by comparing the IQs' external validity; third, we examined possible explanations for these differences; and fourth, we sought possible solutions to the issue of incomparability.

Given an imperfect reliability and results from Floyd et al. (2008), Hagmann-von Arx et al. (2018), and Bünger et al. (2019), we expected to find at least some IQ differences. To examine the possible impact of such differences on external validity, we further determined the IQs' differential relationships with school grades. As IQ is a strong predictor of scholastic achievement and academic success (Deary, Strand, Smith, & Fernandes, 2007; Gygi, Hagmann-von Arx, Schweizer, & Grob, 2017; Roth et al., 2015; Watkins, Lei, & Canivez, 2007), these criteria are typically used for external validation of intelligence tests. In our study, we focused not on the absolute magnitude of relationships between IQ and school grades but rather on possible differences in magnitude of relationships between the FSIQ and school grades and the ScrIQ and school grades. While the former has been studied extensively (see above), to our knowledge, effect sizes of the FSIQ and the ScrIQ have never been compared explicitly, which is what we did in the present study.

After having determined the magnitude and impact of IQ differences, we were interested in possible explanations for these. To achieve this, we explored several possible predictors of IQ differences. These include variables already considered in Hagmann-von Arx et al. (2018) and/or in Bünger et al. (2019), such as IQ level (i.e.,

below average, average, above average) and age, as well as other, not yet examined characteristics of the testee and his or her behavior in the test situation. Characteristics of the composite, such as the number, *g* loadings, and content of subtests involved, might also predict IQ differences, as these characteristics have an influence on the accuracy of composites (see Farmer et al., 2019). Because these characteristics are invariant between individuals, inclusion in quantitative analyses is not possible. Hence, we address them in a descriptive manner only.

As a last step, we explored possible solutions to incomparability. To this end, we examined alternative ways of describing an intelligence test result other than specific IQ scores, aiming to achieve a more reliable and stable estimate. Obvious candidates that were also examined in previous studies (Bünger et al., 2019; Floyd et al., 2008; Hagmann-von Arx et al., 2018) are CIs and nominal categories (e.g., "average" for an IQ between 85 and 115). In all three studies mentioned above, however, CIs were computed solely on the basis of an overall internal consistency, as this reflects the most common use in practice. Results from these studies suggest that using such CIs still does not necessarily lead to satisfactory comparability. As an extension to these previous studies, we therefore varied the reliability coefficients used for the calculation of CIs. As it is known that stability (test–retest reliability) tends to be lower for younger ages (Watkins & Smith, 2013) and toward the tails of the IQ distribution (due to regression to the mean; Campbell & Kenny, 1999), we argue that it is necessary to investigate whether this is also true for internal consistency. If this is the case, using CIs based on separate internal consistency coefficients for age and IQ groups—henceforth called age- and IQ-specific internal consistencies—should lead to higher rates of comparability between IQs compared to using CIs based on the same overall internal consistency for all participants. This assumption is supported by results from Bünger et al. (2019), who found that IQ differences were considerably larger and the percentages of participants with overlapping 95% CIs considerably lower for below-average and above-average IQs compared to average IQs, respectively. A possible influence of age on comparability was not investigated by Floyd et al. (2008) and

Hagmann-von Arx et al. (2018), and as age was not a significant predictor for IQ differences in regression analyses reported in Bünger et al. (2019), no age-differentiated analyses were conducted there either. However, as Bünger et al. (2019) concluded, further analyses with larger age groups are warranted to learn more about IQ comparability across age, which was possible in the present study.

Hence, we investigated comparability across IQ and age for all criteria, and we examined comparability for CIs based on age- and IQ-specific internal consistencies. Moreover, apart from the oversimplification of using a single reliability coefficient for all individuals, using internal consistency as a reliability estimate bears the danger of overestimating reliability, as it misses important an source of measurement error—namely, transient error (see above; Schmidt et al., 2003). This source of error is, however, assessed in test–retest reliability, which is why we also considered CIs based on test–retest reliability coefficients in our study.

Present Study

The primary objective of this study was to investigate the individual-level comparability of different IQs derived from the same test battery. For this purpose, we compared IQs (mostly FSIQ vs. ScrIQ) for participants from the standardization samples of three individually administered test batteries, spanning from early childhood to late adulthood: The Intelligence and Development Scales–2 (IDS-2; Grob & Hagmann-von Arx, 2018b), the German adaptation of the Stanford–Binet Intelligence Scales–Fifth Edition (SB5; Grob, Gygi, & Hagmann-von Arx, 2019a), and the German adaptation of the Reynolds Intellectual Assessment Scales (RIAS; Hagmann-von Arx & Grob, 2014a). Since comparisons of IQs from different intelligence tests were not an aim of this study, we exclusively compared IQs *within* these test batteries. To learn more about the impact of, possible explanations for, and possible solutions to incomparability, secondary objectives of this study were to examine the differential external validity of IQs, to identify predictors for IQ differences, and to investigate how individual-level comparability can be enhanced by varying reliability coefficients used

for the calculation of 95% CIs.

We addressed the following hypotheses and research questions: First, concerning group-level comparisons, we expected (a) that the IQs for each test battery would be highly intercorrelated, and (b) that there would be no significant mean differences between IQs. Second, concerning intraindividual differences, we examined the magnitude of intraindividual differences in IQ points (both overall and across IQ and age). Third, concerning differential external validity, we hypothesized that relationships of school grades with the ScrIQ would be smaller compared to those with the FSIQ. Fourth, concerning possible explanations for IQ differences, we examined whether certain characteristics of the testee (e.g., age) or the testee's behavior in the test situation (e.g., understanding of instructions) were associated with IQ differences. Finally, concerning possible solutions to incomparability, we examined how many participants would achieve comparable intelligence estimates (again both overall and across IQ and age) determined with different criteria (i.e., different 95% CIs and nominal categories). We expected higher comparability for CIs based on age- and IQ-specific internal consistencies and test–retest reliabilities compared to CIs based on one overall internal consistency coefficient.

Method

Participants

The IDS-2 standardization sample consists of 1,672 participants from Switzerland, Germany, and Austria. Complete data on all IQs were available for 1,622 participants (50.9% female; age in years: $M = 12.06$, $SD = 4.40$, range: 5.02–20.97). About one third (31.4%) of participants' mothers had a university degree, 16.5% of participants were bilingual (German and at least one other native language), 7.6% were nonnative speakers (German not their native language), and 3.4% reported having an attention-deficit/hyperactivity disorder (ADHD) or attention-deficit disorder (ADD) diagnosis (hereafter called AD[H]D). For a subsample of 414 individuals (50.7% female; age in years: $M = 12.07$, $SD = 2.59$, range: 5.42–19.37), there were additional

cross-sectional data on school grades.

The SB5 standardization sample consists of 1,829 participants from Switzerland, Germany, Austria, and Liechtenstein. Complete data on all IQs were available for all 1,829 participants (51.4% female; age in years: $M = 23.46$, $SD = 20.02$, range: 4.00–83.96). Around one third (29.4%) of participants—or for children and adolescents, their mothers—had a university degree, 8.8% of participants were bilingual (German and at least one other native language), 7.8% were nonnative speakers (German not their native language), and 2.9% reported having an AD(H)D diagnosis. For a subsample of 249 individuals (47.4% female; age in years: $M = 11.31$, $SD = 2.38$, range: 5.79–17.68), there were additional cross-sectional data on school grades.

The RIAS standardization sample consists of 2,145 participants from Switzerland and Germany. Complete data on all IQs were available for 2,109 participants (49.5% female; age in years: $M = 19.84$, $SD = 20.28$, range: 3.00–99.96). About one fifth (20.7%) of participants—or for children and adolescents, their mothers—had a university degree, and 17.9% of participants were nonnative speakers (German not their native language). For a subsample of 64 individuals, there were additional data on school grades collected 2 to 4 years after the intelligence assessment (51.6% female; age in years at T1: $M = 9.02$, $SD = 1.02$, range: 6.07–11.22, and at T2: $M = 11.41$, $SD = 0.99$, range: 9.00–14.00).

Materials

Intelligence Test Batteries. The IDS-2 assess cognitive (intelligence, executive functions) as well as developmental (psychomotor skills, socioemotional skills, basic skills, and motivation and attitude) in 5- to 20-year-olds with a total of 30 subtests (Grob & Hagmann-von Arx, 2018b; see Table S1 in the supplementary material, `https://osf.io/hfqe5/?view_only=b532b3b095ef417bb4dfa73a7e6c55e3`). The IDS-2 allow for the estimation of three different IQs. The Profile IQ ($\text{PrIQ}_{\text{IDS-2}}$) is based on all 14 subtests that also constitute a profile of the following seven group factors, each estimated by two subtests: Visual Processing, Long-Term Memory,

Processing Speed, Auditory Short-Term Memory, Visual-Spatial Short-Term Memory, Abstract Reasoning, and Verbal Reasoning. The first seven subtests (one per factor) constitute an IQ without a factor profile ($FSIQ_{IDS\text{-}2}$). Additionally, the two subtests with the highest *g* loadings in a confirmatory factor analysis of the first seven subtests—Completing Matrices and Naming Categories—constitute the ScrIQ ($ScrIQ_{IDS\text{-}2}$). Finally, the IDS-2 include a rating of the participation of the testee during testing with 12 questions answered by the test administrator at the end of the intelligence, executive functions, and developmental functions assessments. Here, we used the answers on the intelligence assessment only.

The SB5 are an intelligence test battery for 4- to 83-year-olds that include a total of 10 subtests (Grob et al., 2019a; see Table S1 for descriptions). The following five group factors can be estimated based on one verbal and one nonverbal subtest each: Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial Processing, and Working Memory. Additionally, the five verbal and five nonverbal subtests are used for a Verbal and a Nonverbal IQ. All 10 subtests are used for an FSIQ ($FSIQ_{SB5}$) and the two routing subtests—Nonverbal Fluid Reasoning and Verbal Knowledge—constitute the Abbreviated Battery IQ (ABIQ; hereafter called the $ScrIQ_{SB5}$). Finally, the SB5 include a rating of the participant's understanding of instructions and cooperation in the test situation with one question each answered by the test administrator at the end of the test session.

The RIAS measure verbal and nonverbal intellectual abilities as well as memory with two subtests each (six in total) in 3- to 99-year-olds (Hagmann-von Arx & Grob, 2014a; see Table S1 for descriptions). The two corresponding subtests are used to form a Verbal Intelligence Index, a Nonverbal Intelligence Index, and a Memory Index. All four intelligence subtests are used for an FSIQ ($FSIQ_{RIAS}$). Additionally, one verbal and one nonverbal intelligence subtest—Guess What and Odd-Item Out—constitute the Reynolds Intellectual Screening Test (RIST; hereafter called the $ScrIQ_{RIAS}$).

Having seven available IQs thus enabled five comparisons: three for the IDS-2, one for the SB5, and one for the RIAS. The FSIQs from all three test batteries differ from

each other in terms of number and content of subtests, whereas all ScrIQs consist of one subtest each measuring fluid reasoning and comprehension knowledge (see Table S2 for broad abilities tapped by each IQ). For the $PrIQ_{IDS\text{-}2}$ and $FSIQ_{IDS\text{-}2}$, as well as for the $FSIQ_{RIAS}$ and $ScrIQ_{RIAS}$, only the number of subtests, and not the content, differs, as the corresponding IQs tap the same broad abilities in equal shares. In contrast, for the $PrIQ_{IDS\text{-}2}$/$FSIQ_{IDS\text{-}2}$ and $ScrIQ_{IDS\text{-}2}$, as well as for the $FSIQ_{SB5}$ and $ScrIQ_{SB5}$, content and number of subtests differ.

Participant and Parent Questionnaires. Adolescent and adult participants and/or—for children and adolescents—their parents reported on demographic variables, including age, sex, education (additionally for children and adolescents: education of the parents), native language, and psychological and physical abnormalities (including AD[H]D). In an additional questionnaire, some parents reported their child's school grades in German (instructional language), mathematics, social studies, geography and history (combined), and science from the last two school semesters.

Procedure

Participants were recruited through schools and psychosocial institutions for children and adolescents in Switzerland, Germany, and Austria. For the IDS-2, administration of the whole test battery took between 3 and 4 h and, if necessary, could be split into two sessions no more than 1 week apart. Administration of the intelligence part alone took approximately 1.5 h and was completed within one test session. For the SB5, administration took 1.5 to 2 h and for the RIAS it took around 30 to 40 min. Written consent was obtained from children and adolescents (10 years and older) and/or from their parents (5- to 15-year-olds). The demographic questionnaire was administered at the beginning of the first session. The parental report of school grades was completed at home either within weeks after the session (IDS-2 and SB5) or as part of a follow-up study 2 to 4 years after the intelligence assessment (RIAS). Participants from Switzerland received a gift card of their own choice worth 30 (IDS-2) or 20 (SB5 and RIAS) Swiss francs and participants from Germany and Austria received 25

(IDS-2) or 12 (SB5 and RIAS) euros in cash for participation.

Statistical Analyses

All analyses were conducted in R (R Core Team, 2019). Within each test battery, we first inspected group-level comparability of IQs with Pearson correlations (both uncorrected and corrected for unreliability of both IQs) and paired samples t tests. For individual-level comparability, we then calculated intraindividual absolute differences in IQ points.

To compare the IQs' external validity, we performed linear regressions of school grades on IQ. All grades were transformed into Swiss school grades, ranging from 1 (lowest) to 6 (highest). In our study, we focused on grades in German and mathematics as well as on the grade point average (GPA). The GPA was computed as the average of all reported grades for each participant. As the IQs were expected to be highly correlated, we included them in separate models and compared the resulting R^2s and 95% CIs for standardized regression coefficients (betas). Because the models were not nested, we could not determine the significance of the change in R^2. Instead, following Cumming (2009), we regarded two betas as significantly different from one another if their 95% CIs overlapped to a degree of 50% or less.

To examine possible explanations for IQ differences, we explored several possible predictors, specifically, age, sex, AD(H)D (yes vs. no), native language (monolingual German [reference] vs. bilingual and vs. other native language), IQ level (average [$85 \leq$ IQ ≤ 115, reference] vs. below average [IQ < 85] and vs. above average [IQ > 115]), education of the participant or—for children and adolescents—of their mother (university degree vs. no university degree), participation in the test situation (for IDS-2; age-standardized scores with $M = 10$ and $SD = 3$), cooperation in the test situation (for SB5; yes vs. no/partly), understanding of instructions (for SB5; yes vs. no/partly), and the interaction between IQ and age. To account for the distribution of the dependent variable—IQ differences (see Figure S1)—we used gamma generalized linear models (GLMs) with a log link function. Following suggestions from Gelman

(2008), we standardized all predictor variables by dividing by 2 *SD*s. We deemed an effect significant if both the overall model (determined with a likelihood ratio test) and the predictor were significant at an alpha level of .05. To illustrate the variation of IQ differences across IQ and age, we compared the resulting difference scores across IQ (including six IQ groups: < 70, 70–84, 85–99, 100–114, 115–129, ≥ 130; see Figure 1) and across age (including different age groups depending on the test battery; see Figure 2). The IQ groups were based on the IQ with the largest number of subtests for each test battery (i.e., the $\text{PrIQ}_{\text{IDS-2}}$, FSIQ_{SB5}, and $\text{FSIQ}_{\text{RIAS}}$). The same IQs were used for the predictor of IQ level in regression analyses.

To explore possible solutions to incomparability, we computed 95% CIs using the standard error of estimate (SEE) together with the estimated true score (Lord & Novick, 1968; see also Dudek, 1979). For each test battery, we then calculated the percentage of participants for whom the 95% CIs for the IQs overlapped. We varied the reliability coefficients used for the calculation of 95% CIs to investigate their influence on individual-level comparability. The 95% CIs were based on overall internal consistencies (95CI; for IDS-2 and RIAS: Cronbach's alphas and for SB5: split-half reliabilities), age-specific internal consistencies (95CI_{age}; see Table S9 for age groups), and test–retest reliabilities (95CI_{rtt}) obtained from the test manuals (Grob, Gygi, & Hagmann-von Arx, 2019b; Grob & Hagmann-von Arx, 2018a; Hagmann-von Arx & Grob, 2014b). Additionally, we calculated 95% CIs based on age- and IQ-specific internal consistencies according to the manuals using a formula provided by Lienert and Raatz (1998, p. 330; $95\text{CI}_{\text{ageIQ}}$; e.g., for 5- to 6-year-olds with IQ < 85; see Table 4 for IQ and age groups). Finally, we investigated the comparability of the IQs' corresponding nominal categories (NomIQ; < 70 = "lower extreme," 70–84 = "below average," 85–115 = "average," 116–130 = "above average," > 130 = "upper extreme"; see also Grob, Meyer, & Hagmann-von Arx, 2013) as well as the comparability of the 95% CIs with overall internal consistencies in nominal categories (NomCI; e.g., average to above average for an interval of 112 to 120). For each of these six resulting criteria—95CI, 95CI_{age}, $95\text{CI}_{\text{ageIQ}}$, 95CI_{rtt}, NomIQ, and NomCI—two IQs were deemed

comparable on an individual level if their intervals overlapped. Just as for IQ differences, we compared the percentages of participants with overlapping intervals across IQ and age using the same groups.

Results

Group-Level Analyses

The seven IQs considered were normally distributed; their means were close to 100 (99.53 to 100.11) and standard deviations close to 15 (14.49 to 15.11, see Table 1). The $FSIQ_{IDS\text{-}2}$ had the narrowest range with 55 to 142, and the $ScrIQ_{RIAS}$ had the widest range with 40 to 160. We compared the IQs within each test battery using t tests and Pearson correlations and found very small mean differences that were non-significant in all but one case (d = -0.002 for the $FSIQ_{IDS\text{-}2}$ vs. the $ScrIQ_{IDS\text{-}2}$ to $d = 0.031$ for the the $FSIQ_{RIAS}$ vs. the $ScrIQ_{RIAS}$), the latter being significant, $t(2108) = 3.73$, $p < .001$. Intercorrelations both uncorrected and corrected for unreliability of both IQs were all significant and high to very high ($r = .76$ for the $FSIQ_{SB5}$ and the $ScrIQ_{SB5}$ to $r = .95$ for the $PrIQ_{IDS\text{-}2}$ and the $FSIQ_{IDS\text{-}2}$, and $r_{corr} = .77$ for the $FSIQ_{SB5}$ and the $ScrIQ_{SB5}$ to $r_{corr} = .99$ for the $FSIQ_{RIAS}$ and the $ScrIQ_{RIAS}$, all with $p < .001$).

Intraindividual Differences

The mean (and median) intraindividual absolute differences ranged between 3.68 ($Mdn = 3$) IQ points for the $PrIQ_{IDS\text{-}2}$ versus the $FSIQ_{IDS\text{-}2}$ and 8.12 ($Mdn = 7$) IQ points for the $FSIQ_{SB5}$ versus the $ScrIQ_{SB5}$, with ranges between 0 and 20 ($PrIQ_{IDS\text{-}2}$ vs. $FSIQ_{IDS\text{-}2}$ and $FSIQ_{RIAS}$ vs. $ScrIQ_{RIAS}$) and 0 and 39 IQ points ($FSIQ_{IDS\text{-}2}$ vs. $ScrIQ_{IDS\text{-}2}$; see Table 2). The relative differences were normally distributed around 0 (see Figure S2). Absolute differences across IQ groups and age are displayed in Figures 1 and 2, respectively (see also Table S3). For most comparisons, differences tended to increase with higher IQs and for the $PrIQ_{IDS\text{-}2}$ versus the $FSIQ_{IDS\text{-}2}$ and the $FSIQ_{RIAS}$ versus the $ScrIQ_{RIAS}$, they tended to decrease with lower IQs. Regarding age, differences were lowest for middle childhood for the $FSIQ_{SB5}$ versus the $ScrIQ_{SB5}$, but

highest for the same age period for the $FSIQ_{RIAS}$ versus the $ScrIQ_{RIAS}$. Otherwise, differences showed little variation across age.

Differential Relationships With School Grades

To compare the IQs' external validity, we investigated their differential relationships with school grades in German and mathematics, and with the GPA. Comparisons of 95% CIs for the betas revealed that the relationship with the FSIQ was significantly higher than that with the ScrIQ only for the SB5 and mathematics (see Figure S3 and Table S4).

In a post hoc analysis, we repeated the external validity analyses for subsamples with small (below median) and large (above median) IQ differences to see how incomparability might affect external validity (see Figure 3 and Table S5). For individuals with small IQ differences, we found small to medium relationships that were all highly significant ($\beta = .29$ for the $FSIQ_{IDS\text{-}2}$ and German to $\beta = .52$ for the $ScrIQ_{RIAS}$ and German, all with $p < .001$), and there were no significant differences in betas between IQs. For individuals with large IQ differences, however, betas were still significant for the IDS-2 and SB5 ($\beta = .18$, $p = .008$ for the $ScrIQ_{IDS\text{-}2}$ and mathematics to $\beta = .46$, $p < .001$ for $FSIQ_{SB5}$ and mathematics), but lower for the SB5 and no longer significant for the RIAS ($\beta = .01$, $p = .965$ for the $FSIQ_{RIAS}$ and mathematics to $\beta = .12$, $p = .483$ for the $ScrIQ_{RIAS}$ and mathematics). For the IDS-2 and SB5, relationships with the ScrIQ were also consistently smaller compared to those with the FSIQ and the PrIQ, although for both, this difference in betas was only significant for mathematics (see Figure 3).

Possible Predictors of IQ Differences

Next, we investigated possible predictors of IQ differences. Only the models for the comparisons of the $PrIQ_{IDS\text{-}2}$ versus the $ScrIQ_{IDS\text{-}2}$, the $FSIQ_{SB5}$ versus the $ScrIQ_{SB5}$, and the $FSIQ_{RIAS}$ versus the $ScrIQ_{RIAS}$ were significant. Therein, IQ level and age and/or their interaction were the only consistent predictors (see Table 3, see Table S6 for results for all comparisons). Larger differences occurred for younger

individuals for two comparisons: for individuals with a below-average IQ for two comparisons, and for individuals with an above-average IQ for one comparison. Finally, there was a significant interaction effect for age and below-average IQ for one comparison and for age and above-average IQ for another comparison (see supplementary material for a detailed description of results).

Comparability Using Different Criteria

Table 4 shows the reliabilities and widths of the corresponding 95% CIs for all seven IQs. The width of the 95% CIs based on overall internal consistencies ranged between 6 ($FSIQ_{SB5}$) and 15 ($ScrIQ_{RIAS}$) IQ points. Those based on age-specific internal consistencies and test–retest reliabilities were considerably larger, and those based on age- and IQ-specific internal consistencies reached up to 30 IQ points for some combinations of IQ > 115 and different age groups. The lowest age- and IQ-specific internal consistencies, resulting in the largest CIs, were found exclusively in groups with IQ > 115 and did not coincide with the lowest sample sizes for any of the IQs.

The percentage of participants with comparable IQs (i.e., overlapping intervals) varied considerably across the different criteria and across IQ and age groups (see Figure 4 and Tables S7 to S12). With the 95CI criterion, overall comparability was between 60.5% and 98.7%. Across IQ groups it ranged between 27.8% and 99.6% and across age groups between 50.7% and 100%.

The overall comparability was lowest for the NomIQ (69.9% to 87.5%) and the 95CI (60.5% to 98.7%) criteria and highest for the $95CI_{rtt}$ (94.3% to 100.0%) and the $95CI_{ageIQ}$ (96.7% to 99.9%) criteria. The same pattern was evident across IQ and age groups, with the lowest comparability for the NomIQ and the 95CI and highest comparability for the $95CI_{rtt}$ and the $95CI_{ageIQ}$. In general, comparability was lowest for the comparison of the $FSIQ_{SB5}$ versus the $ScrIQ_{SB5}$ and highest for the comparison of the $FSIQ_{RIAS}$ versus the $ScrIQ_{RIAS}$.

Discussion

The primary objective of this study was to investigate the individual-level comparability and external validity of different IQs derived from the same test battery. As expected, all IQs were highly intercorrelated and—with one exception—there were no significant mean differences. Despite this high correspondence on the group level, individual-level comparability was not always satisfactory. Intraindividual absolute differences reached up to 39 IQ points and tended to be larger for above-average IQ and younger ages. However, with respect to external validity, the PrIQ and the FSIQ explained more variance in school grades compared to the ScrIQ only for individuals with large IQ differences and only for the IDS-2 and the SB5, with significant differences only for mathematics. Regarding possible explanations, IQ level and age, and/or their interaction, were the only consistent predictors of IQ differences. Finally, regarding possible solutions, comparability varied considerably across criteria and again across both IQ and age within a comparison and between comparisons. While comparability for the NomIQ and 95CI was often unsatisfactory, it was very high for the $95CI_{rtt}$ and $95CI_{ageIQ}$.

Group-Level Comparability and Intraindividual Differences

On the group level, all IQs within each test battery were highly comparable, with the exception of the $FSIQ_{RIAS}$ and $ScrIQ_{RIAS}$, where we found a significant mean difference despite a very high correlation. However, the effect size was very small, suggesting the effect is negligible. Despite high comparability on the group level, intraindividual absolute differences between IQs varied considerably, from 0 to more than 2.5 *SD*s, with means between a fourth and over half a standard deviation, depending on the comparison. It is important to note that there were no systematic differences in one direction (one IQ being systematically higher than the other) as the relative differences were normally distributed around 0 for all comparisons. The mean IQ differences were slightly lower than those found in previous studies investigating individual-level comparability of FSIQs between test batteries (Bünger et al., 2019;

Floyd et al., 2008; Hagmann-von Arx et al., 2018). Still, the size of the differences seems remarkable, given that the subtests for the ScrIQ are fully included in the FSIQ and the PrIQ, and the subtests for the FSIQ are fully included in the PrIQ, that transient error is kept to a minimum, and that between-battery variance is ruled out completely.

Differential External Validity

Analyses on differential external validity revealed that the PrIQ and the FSIQ had significantly stronger relationships with school grades compared to the ScrIQ only for participants with large IQ differences and only for the IDS-2 and the SB5 and mathematics. These two comparisons—$PrIQ_{IDS\text{-}2}$ versus $ScrIQ_{IDS\text{-}2}$ and $FSIQ_{SB5}$ versus $ScrIQ_{SB5}$—also featured the largest IQ differences. Further, there were no differences in relationships between the $PrIQ_{IDS\text{-}2}$ and the $FSIQ_{IDS\text{-}2}$. Apparently, the ScrIQs miss aspects of intelligence that are contained in the PrIQ and FSIQs that are especially important for mathematical achievement. For the IDS-2, this probably concerns additional working memory aspects (tapped at least in part by the two/four short-term memory subtests) and visual-spatial skills (tapped by the [two] Visual Processing and the [two] Visual-Spatial Short-Term Memory subtests; see Table S1) tapped by the FSIQ/PrIQ but not the ScrIQ, as these abilities are known to be especially related to mathematical achievement (e.g., Bull & Lee, 2014; Kahl, Grob, Segerer, & Möhring, 2019; McCrink & Opfer, 2014). For the SB5, displaying the largest difference between the FSIQ and ScrIQ, the incremental validity of the FSIQ is probably mostly due to the two Quantitative Knowledge subtests, in addition to the two Working Memory and two Visual-Spatial Processing subtests.

Moreover, relationships were smaller for individuals with large compared to small IQ differences for the SB5 and the RIAS, to the point that for the RIAS, they were no longer significant for individuals with large IQ differences. Although the sample sizes were small for the RIAS here, and therefore findings have to be interpreted with caution, it is still important to take a closer look at this particular finding, as it is consistent and based on a longitudinal analysis. This finding might indicate that for

individuals with larger discrepancies between different IQs, it is possible that none of these IQs predict school grades in the long run.

From these findings we conclude that an IQ based on more subtests is not necessarily a better predictor for school grades compared to one based on fewer subtests, especially for individuals with low IQ differences. We also conclude that larger IQ differences do have consequences for external validity, as the IQs for which larger intraindividual differences occurred were also the ones with larger disparities in relationships with school grades, and as relationships tended to be lower in general for individuals with high IQ differences. For these individuals, relationships with the ScrIQ also tended to be lower compared to those with the FSIQ and the PrIQ, especially for mathematics. In these cases, differences in content seem to be more important than differences in the number of subtests per se.

Possible Explanations

If the IQ differences we found are not entirely due to random error, they most likely stem from (stable) characteristics of the testee and/or of the composites themselves. Our results suggest that the first possibility (i.e., characteristics of the testee) is likely part of the answer, as IQ differences varied across IQ and age, and those two and/or their interaction were the only systematic predictors in regression analyses. In this respect, our results are in line with Hagmann-von Arx et al. (2018) and Bünger et al. (2019), where IQ differences were larger for younger participants for some comparisons and larger at the tails of the IQ distribution for some (Hagmann-von Arx et al., 2018) or most (Bünger et al., 2019) comparisons as well. In our study (and not investigated in previous studies) there were also significant interaction effects between IQ level and age, indicating these two variables should be considered in conjunction with each other when calculating reliability coefficients. Finally, the included predictors explained a significant amount of variance for only three of the five comparisons. It is likely that other variables that could not be sufficiently considered in the present study contribute to systematic variance in IQ differences, for example (achievement)

motivation, attention span, or alertness. Thus, there are two characteristics of the testee (i.e., IQ and age) that explain some of the variance in IQ differences. The second source of systematic variability introduced above (i.e., characteristics of the composites) therefore likely played a role as well. Three such characteristics are number, *g* loadings, and content of subtests going into the composites. As mentioned above, Farmer et al. (2019) showed that the most accurate composites are those derived from numerous (12 to 13) diverse and highly *g*-loaded subtests, where high *g* loadings are more important compared to heterogeneity. Their results also suggest that composites based on four subtests might be accurate enough, but fewer than four subtests resulted in substantial losses of accuracy. In line with common practice, the ScrIQs from the three test batteries included in our study are all composed of only two subtests. Further, although all three ScrIQs fulfill the heterogeneity criterion with the two subtests representing two different broad abilities (namely, fluid reasoning and comprehension knowledge), only the subtests for the $\text{ScrIQ}_{\text{IDS-2}}$ were chosen based on the highest *g* loadings. The $\text{ScrIQ}_{\text{SB5}}$ is composed of the subtests with the lowest (Nonverbal Fluid Reasoning) and third lowest (Verbal Knowledge) *g* loading (Grob et al., 2019b), which might at least partly explain the larger differences we found for the SB5 compared to the IDS-2 and the RIAS.

Subtest content may also explain some of the variance in individual-level comparability between the different comparisons. In this regard, it is especially interesting to compare the comparisons of the $\text{PrIQ}_{\text{IDS-2}}$ versus the $\text{FSIQ}_{\text{IDS-2}}$ and the $\text{FSIQ}_{\text{RIAS}}$ versus the $\text{ScrIQ}_{\text{RIAS}}$. Both comparisons have the same degree of overlap in content (100%, see Table 4) and the same ratio of subtests (2:1) but different absolute numbers of subtests (4 and 2 vs. 14 and 7) and different numbers of broad abilities tapped (2 vs. 7). Differences for the $\text{PrIQ}_{\text{IDS-2}}$ versus the $\text{FSIQ}_{\text{IDS-2}}$ are slightly lower than for the $\text{FSIQ}_{\text{RIAS}}$ versus the $\text{ScrIQ}_{\text{RIAS}}$, but both are considerably lower compared to the other comparisons.

From these considerations we conclude that larger overlap in content and high *g* loadings seem to be more important than the sheer number of subtests for high

individual-level comparability. However, as our set of comparisons is very limited, these findings clearly need replication, ideally with comparisons of composites systematically varied in content, g loadings, and number of subtests.

Possible Solutions

We explored several alternatives to specific IQ scores—nominal categories and 95% CIs based on different reliability coefficients—with the aim of achieving a more dependable intelligence estimate. Results on percentages of participants with overlapping 95% CIs or nominal IQs reflect results on IQ differences in that they varied both between the different comparisons and across IQ and age. Although all investigated criteria consider unreliability in one way or another, comparability still tended to be lower at younger ages and toward the tails of the IQ distribution.

Furthermore, comparability varied considerably between the different criteria. Although the overall percentages of participants with overlap of the 95CI and the NomIQ tended to be higher compared to those found in previous studies on between-battery comparisons (Bünger et al., 2019; Hagmann-von Arx et al., 2018), they were still unsatisfactory. Especially when calculated separately for IQ and age groups, the percentage of participants with comparable IQs was sometimes very low, down to 27.8%, which is comparable to results from Bünger et al. (2019). Rates of comparability were higher for the $95CI_{age}$ and the NomCI, criteria but the highest rates were achieved with the $95CI_{rtt}$ or the $95CI_{ageIQ}$ criteria. This is to be expected, given that the intervals were also often widest for these criteria compared to the other criteria. Which of the two—$95CI_{rtt}$ or $95CI_{ageIQ}$—provides a better trade-off between comparability and precision (in terms of interval width) is difficult to pin down as this varies across IQs and across IQ comparisons. At this point, it is important to note that we had to rely on fairly rough groups for IQ (< 85, 85–115, and > 115) and for age in adulthood (e.g., age 30–59 years for the SB5 and age 21–59 years for the RIAS). Additionally, group sizes varied considerably and were sometimes very low (IDS-2: $n = 31$ to $n = 352$; SB5: $n = 15$ to $n = 222$; RIAS: $n = 23$ to $n = 175$). The comparability versus precision trade-off

could probably be improved for the $95CI_{ageIQ}$ if larger, more fine-graded groups were considered, which would necessitate sampling more participants of diverse ages at the tails of the IQ distribution. Finally, both internal consistency and test–retest reliability miss certain kinds of measurement error. While internal consistency does not consider transient error (see above), test–retest reliability does not consider specific factor error (i.e., errors due to individual interpretation of items; Schmidt et al., 2003). Therefore, although we recommend both test–retest reliability and age- and IQ-specific internal consistencies over one overall internal consistency as a basis for CIs, other approaches may be even more beneficial. The coefficient of equivalence and stability (CES; Cronbach, 1947), for example, combines the advantages of internal consistency and test–retest reliability and considers both specific factor error and transient error, making it likely to be the best candidate as a basis for CIs. As CES requires the administration of two parallel test forms on two different measurement occasions, we were not able to consider this reliability coefficient in our study.

Finally, we would like to emphasize that more accurate confidence intervals can be only part of the solution to incomparability, mainly as a means for practitioners to deal with incomparability of results from existing intelligence tests. Given the substantial differences we found, the consequences they have for validity, and the large intervals needed to achieve satisfactory individual-level comparability, the long-term goal must be to create more accurate intelligence measures. To achieve a higher individual-level comparability, it might be necessary to question our current understanding of general intelligence and to refrain from multidimensional measures (i.e., subtests intended to measure both general intelligence and a broad ability; see also Beaujean & Benson, 2019). Instead, test developers could try to create unidimensional measures of specific broad abilities—including clean fluid reasoning measures as a substitute for currently used general intelligence composites—with a firmer theoretical and neurological basis (e.g., Beaujean & Benson, 2019; Kovacs & Conway, 2019).

Implications

Our findings have implications for the construction, validation, and application of intelligence tests. For test construction and application, these findings raise awareness that choosing the subtests with the highest *g* loadings (i.e., those most representative for the whole scale) for a short form does not necessarily result in comparable results to those for the full test battery. However, it is certainly better than choosing subtests with lower *g* loadings (see also Farmer et al., 2019).

Moreover, our results indicate that in terms of both individual-level comparability and external validity there are no large gains between the 7- and 14-subtest composites (the FSIQ and the PrIQ, respectively) for the IDS-2. In line with results from Farmer et al. (2019), this suggests a diminishing marginal utility of additional subtests—especially if they do not introduce other broad abilities—from a certain number of subtests on.

Our finding of the importance of content for comparability suggests that when constructing a short form or short test—under the current conception of intelligence as a construct comprising different factors or broad abilities (McGrew, 2009; reservations discussed above notwithstanding)—it is likely better to tap more broad abilities with one subtest each, at the cost of having no factor profile (e.g., four subtests that capture four broad abilities), than to measure less broad abilities to have a factor profile (e.g., four subtests that together capture two broad abilities). The former should also render a more accurate *g* in the sense of Jensen and Weng's (1994) definition by providing a more diverse set of subtests regarding content.

Finally, on the basis of our results, we argue against using one internal consistency coefficient derived from the whole sample for the calculation of CIs. Instead, we recommend the use of test–retest reliabilities, age- and IQ-specific internal consistencies or, probably even better, the CES (Schmidt et al., 2003). The additional resources spent on the construction and application of a parallel test form would likely be more than compensated for by more accurate reliability estimates—and with this, more accurate CIs—and even by the possibility to provide practitioners with a test battery they could administer twice to the same testee without the interference of learning

effects. Ideally, but of course difficult to implement in practice, the test–retest sample should also be large enough to permit at least a rough division into IQ and age groups to enable the use of age- and IQ-specific CESs for the calculation of CIs.

An important implication for practitioners is not to use specific IQ scores for interpretation or communication of test results. Indeed, in line with Bünger et al. (2019), our results show that even the 95% CI might not necessarily be valid enough for clinical interpretation, but it is certainly more appropriate than a specific IQ score. As has been done before (Bünger et al., 2019), we therefore again call for a paradigm shift away from specific IQ scores toward intervals that consider the unreliability of IQ measures in clinical interpretation. Instead of requiring an IQ score to fall above or below a certain threshold, the upper and lower levels of the 95% CI should be considered.

Many practitioners base decisions about the need for further testing on the screening result (Thompson et al., 2004). If the result is atypical, they would consider further testing; if not, they would refrain from further testing. However, our results demonstrate that the differences between the FSIQ and the ScrIQ are largest especially in those ranges where most clinical questions arise—namely, at the tails of the IQ distribution. This is true even if 95% CIs are based on the expected true score, thus accounting for regression to the mean. To avoid the risk of missing diagnostically meaningful information, when time constraints are high, we suggest using a short test of at least four subtests (see Farmer et al., 2019) instead of a screening with two subtests, also for screening purposes. There is another context, gaining importance in many Western countries, in which very short measures should be avoided: for testees with low familiarity with (standardized) testing or test content as well as with difficulties in understanding task instructions. Following insights from dynamic testing (Beckmann, 2014; Beckmann & Dobat, 2000; Cho & Compton, 2015; Guthke & Wiedl, 1996), test performance of such testees increases in predictive validity with increasing time spent with the tasks. For example, it was shown that in a test–retest design, performance in the posttest was a better predictor for scholastic achievement compared to performance

in the pretest, especially for disadvantaged children (Guthke & Wiedl, 1996). The use of a screening instrument thus bears the risk of underestimating an individual's true intellectual potential especially in these contexts.

Finally, IQ differences are linked to the prediction of school grades. For individuals with higher IQ differences, relationships with school grades tended to be lower in general, and especially for the ScrIQ. In the long run, IQ might not even be predictive at all for school grades for individuals with high IQ differences. It is therefore important to identify these individuals, for example, through multiple testing, and to be aware of the possibility of reduced reliability and (external) validity of IQ in these cases.

Future research should determine to what extent the present results are applicable to factor index scores as well. If two subtests are likely not enough for a general intelligence measure, this should be even less appropriate for a factor index score, given the small unique variance over and above g such factors already capture (e.g., Canivez, Watkins, & Dombrowski, 2016; Dombrowski, 2014). We also advocate the use of individual-level comparisons in addition to group-level analyses for validation of a test procedure intended for individual diagnostics. More research is needed to further investigate characteristics of both the testee and the test itself that lead to individual-level comparability (or incomparability) of IQs. Moreover, further research is needed to determine which reliability coefficient(s) should be used for the calculation of CIs to achieve the best trade-off between individual-level comparability and precision, with age- and IQ-specific CESs being a promising candidate. Finally, in addition to internal and structural validation, a greater emphasis should be placed on external validation, but also on diagnostic and treatment utility, of test scores to determine their usefulness as a diagnostic instrument in practice.

Strengths and Limitations

For the first time, we investigated group- and individual-level comparability of IQs within the same test battery for a set of three test batteries based on large, representative samples covering a large age span from early childhood to late adulthood.

In comparing IQs within test batteries, we were able to eliminate all kinds of variance between test batteries or test situations, leaving characteristics of the testee and the test itself as the only systematic sources of variance.

A limitation of this study is that we could form only broad IQ groups for age- and IQ-specific 95% CIs (below average, average, above average) due to small sample sizes within age groups. Greater oversampling of participants of different ages at the tails of the IQ distribution is needed to achieve more fine-graded groups and with this to ensure reliability and validity at the extremes.

Furthermore, we used school grades as a single criterion of external validity. Although school grades are strongly related to IQ (Roth et al., 2015), future research should consider differential relationships of IQs based on different numbers of subtests with additional criteria for scholastic achievement, such as scholastic aptitude tests or teacher ratings of school performance, as well as with criteria that are also valid for adults, for example, educational attainment or occupational success.

Finally, we could include only a limited number of test batteries and composites in our study. Systematic comparisons of the kind performed in Farmer et al. (2019)—comparisons of composites systematically varied in characteristics such as number, *g* loadings, and content of subtests—but on an individual level and within multiple test batteries are needed to further clarify the number and nature of subtests necessary to achieve an accurate measure of general intelligence.

Conclusion

Our findings raise awareness of the limitations of IQ screenings as a means to get a first impression of an individual's intellectual potential. Despite high comparability on the group level, individual-level comparability of IQs derived from the same test battery was often unsatisfactory. We therefore advocate acknowledging a lower reliability of IQs to achieve more accurate intelligence assessments. One step in that direction would be to refrain from using internal consistencies and to instead use test–retest reliabilities or, probably even better, the CES (Cronbach, 1947) as a basis for CIs. The systematic

effects of IQ level and age on IQ differences we found also suggest that reliabilities should be computed separately for age and IQ groups. Most importantly, our results demonstrate that the interpretation of specific IQ scores should be avoided. However, despite limited comparability with the FSIQ, we found that ScrIQs did not necessarily display less external validity. But IQs in general, and especially ScrIQs, tended to be worse predictors of school grades, especially in mathematics, for individuals with large intraindividual IQ differences.

To conclude, our results point to substantial intraindividual IQ differences that have consequences for external validity and are at least in part explained by IQ level and age. Our results demonstrate that a focus on CIs based on reasonable reliability coefficients can be part of the solution to incomparability. Yet, further research is needed to learn more about the number and kind of subtests necessary to achieve an accurate measurement of general intelligence on the individual level.

References

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll Theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 185–202). New York, NY: Guilford Press.

Batty, G. D., Gale, C. R., Tynelius, P., Deary, I. J., & Rasmussen, F. (2009). IQ in early adulthood, socioeconomic position, and unintentional injury mortality by middle age: A cohort study of more than 1 million Swedish men. *American Journal of Epidemiology*, *169*, 606–615. doi: 10.1093/aje/kwn381

Beaujean, A. A., & Benson, N. F. (2019). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology*, *23*, 126–137. doi: 10.1007/s40688-018-0182-1

Beckmann, J. F. (2014). The umbrella that is too wide and yet too small: Why dynamic testing has still not delivered on the promise that was never made. *Journal of Cognitive Education and Psychology*, *13*, 308–323. doi: 10.1891/1945-8959.13.3.308

Beckmann, J. F., & Dobat, H. (2000). Zur Validierung der Diagnostik intellektueller Lernfähigkeit [On the validation of intellectual learning ability diagnostics]. *Zeitschrift für Pädagogische Psychologie*, *14*, 96–105. doi: 10.1024//1010-0652.14.23.96

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*, 36–41. doi: 10.1111/cdep.12059

Bünger, A., Grieder, S., Schweizer, F., & Grob, A. (2019). *The comparability of intelligence test results: Group- and individual-level comparisons of seven intelligence tests.* Manuscript in preparation.

Campbell, D., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York, NY: Guilford Press.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the

Wechsler Intelligence Scale for Children-Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, *28*, 975–986. doi: 10.1037/pas0000238

Cho, E., & Compton, D. L. (2015). Construct and incremental validity of dynamic assessment of decoding within and across domains. *Learning and Individual Differences*, *37*, 183–196. doi: 10.1016/j.lindif.2014.10.004

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika*, *12*, 1–16. doi: 10.1007/BF02289289

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, *28*, 205–220. doi: 10.1002/sim.3471

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21. doi: 10.1016/j.intell.2006.02.001

Dombrowski, S. C. (2014). Investigating the structure of the WJ-III Cognitive in early school age through two exploratory bifactor analysis procedures. *Journal of Psychoeducational Assessment*, *32*, 483–494. doi: 10.1177/0734282914530838

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335–337. doi: 10.1037/0033-2909.86.2.335

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. doi: 10.1037//1082-989X.1.2.170

Farmer, R., Floyd, R., Berlin, K. S., & Reynolds, M. (2019). How can general intelligence composites more accurately index psychometric g and what might be good enough? *Contemporary School Psychology*, Advance online publication. doi: 10.1007/s40688-019-00244-1

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice*, *39*, 414–423. doi: 10.1037/0735-7028.39.4.414

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations.

Statistics in Medicine, *27*, 2865–2873. doi: 10.1002/sim.3107

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence*, *24*, 13–23. doi: 10.1016/S0160-2896(97)90011-8

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology*, *86*, 174–199. doi: 10.1037/0022-3514.86.1.174

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science*, *13*, 1–4. doi: 10.1111/j.0963-7214.2004.01301001.x

Grob, A., Gygi, J. T., & Hagmann-von Arx, P. (2019a). *The Stanford-Binet Intelligence Scales–Fifth Edition (SB5)–German adaptation.* Bern, Switzerland: Hogrefe.

Grob, A., Gygi, J. T., & Hagmann-von Arx, P. (2019b). *The Stanford-Binet Intelligence Scales–Fifth Edition (SB5)–German adaptation. Test manual.* Bern, Switzerland: Hogrefe.

Grob, A., & Hagmann-von Arx, P. (2018a). *Intelligence and Development Scales–2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche.* [Intelligence and Development Scales for children and adolescents]. Bern, Switzerland: Hogrefe.

Grob, A., & Hagmann-von Arx, P. (2018b). *Intelligence and Development Scales–2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche. Manual zu Theorie, Interpretation und Gütekriterien.* [Intelligence and Development Scales for children and adolescents. Manual on theory, interpretation, and psychometric criteria]. Bern, Switzerland: Hogrefe.

Grob, A., Meyer, C. S., & Hagmann-von Arx, P. (2013). *Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5-10 Jahren. Manual.* [Intelligence and Development Scales for Children from 5-10 years of age. Manual]. Bern, Switzerland: Hans Huber.

Guthke, J., & Wiedl, K. H. (1996). *Dynamisches Testen: Zur Psychodiagnostik der*

intraindividuellen Variabilität. [Dynamic testing: On the psychodiagnostics of intraindividual variability]. Göttingen, Germany: Hogrefe.

Gygi, J. T., Hagmann-von Arx, P., Schweizer, F., & Grob, A. (2017). The predictive validity of four intelligence tests for school grades: A small sample longitudinal study. *Frontiers in Psychology*, *8*. doi: 10.3389/fpsyg.2017.00375

Hagmann-von Arx, P., & Grob, A. (2014a). *Reynolds Intellectual Assessment Scales and Screening (RIAS)^{TM}. German adaptation of the Reynolds Intellectual Assessment Scales (RIAS)^{TM} & the Reynolds Intellectual Screening Test (RIST)^{TM} from Cecil R. Reynolds and Randy W. Kamphaus.* Bern, Switzerland: Hans Huber.

Hagmann-von Arx, P., & Grob, A. (2014b). *Reynolds Intellectual Assessment Scales and Screening (RIAS)^{TM}. German adaptation of the Reynolds Intellectual Assessment Scales (RIAS)^{TM} & the Reynolds Intellectual Screening Test (RIST)^{TM} from Cecil R. Reynolds and Randy W. Kamphaus. Test manual.* Bern, Switzerland: Hans Huber.

Hagmann-von Arx, P., Lemola, S., & Grob, A. (2018). Does IQ = IQ? Comparability of intelligence test scores in typically developing children. *Assessment*, *25*, 691–701. doi: 10.1177/1073191116662911

Jensen, A. R., & Weng, L.-J. (1994). What is a good g? *Intelligence*, *18*, 231–258. doi: 10.1016/0160-2896(94)90029-9

Kahl, T., Grob, A., Segerer, R., & Möhring, W. (2019). Executive functions and visual-spatial skills predict mathematical achievement: Asymmetrical associations across age. *Psychological Research*, Advance online publication. doi: 10.1007/s00426-019-01249-4

Kovacs, K., & Conway, A. R. A. (2019). A unified cognitive/differential approach to human intelligence: Implications for IQ testing. *Journal of Applied Research in Memory and Cognition*, *8*, 255–272. doi: 10.1016/j.jarmac.2019.05.003

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6th ed.). Weinheim, Germany: Beltz.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores.* Oxford,

England: Addison-Wesley.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General intelligence,' objectively determined and measured". *Journal of Personality and Social Psychology*, *86*, 96–111. doi: 10.1037/0022-3514.86.1.96

McCrink, K., & Opfer, J. E. (2014). Development of spatial-numerical associations. *Psychological Science*, *23*, 439–445.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–179). New York, NY: Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1–10. doi: 10.1016/j.intell.2008.08.004

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. doi: 10.1016/j.intell.2015.09.002

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*, 206–224. doi: 10.1037/1082-989X.8.2.206

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292. doi: 10.2307/1412107

Thompson, A. P., LoBello, S. G., Atkinson, L., Chisholm, V., & Ryan, J. J. (2004). Brief intelligence testing in Australia, Canada, the United Kingdom, and the United States. *Professional Psychology: Research and Practice*, *35*, 286–290. doi: 10.1037/0735-7028.35.3.286

Watkins, M. W., Lei, P.-W., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence*, *35*, 59–68. doi: 10.1016/j.intell.2006.04.005

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children–Fourth Edition. *Psychological Assessment*, *25*, 477–483. doi: 10.1037/a0031653

Table 1

Descriptive Statistics for IQs and Group-Level Comparisons with Paired-Samples t *Tests and Pearson Correlations*

IQ	*M*	*SD*	Range	Skewness	Kurtosis	*t*		Cohen's *d*	*r*		r_{corr}	
$PrIQ_{IDS\text{-}2}$	100.04	14.70	55–145	-0.49	0.65	-0.61		-0.01	.95	***	.98	***
$FSIQ_{IDS\text{-}2}$	100.11	14.79	55–142	-0.44	0.48	-0.12		-0.00	.82	***	.86	***
$ScrIQ_{IDS\text{-}2}$	100.08	15.11	55–144	-0.30	0.10	-0.12		-0.00	.77	***	.80	***
$FSIQ_{SB5}$	99.96	14.80	55–145	-0.02	0.22	-0.18		-0.00	.76	***	.77	***
$ScrIQ_{SB5}$	99.92	14.95	55–145	-0.06	0.00							
$FSIQ_{RIAS}$	99.53	14.77	45–158	-0.49	0.91	3.73	***	0.03	.93	***	.99	***
$ScrIQ_{RIAS}$	99.98	14.49	40–160	-0.79	1.63							

Note. IDS-2: N = 1,622; SB5: N = 1,829; RIAS: N = 2,109. The last four columns refer to the comparison between the respective IQ and the one in the row below it (for the $ScrIQ_{IDS\text{-}2}$: with the $PrIQ_{IDS\text{-}2}$). Cohen's d was calculated using the formula from Dunlap, Cortina, Vaslow, and Burke (1996) for paired samples. PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ; IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; M = mean; SD = standard deviation, r_{corr} = Pearson correlation corrected for unreliability of both IQs.

***$p < .001$

Table 2

Intraindividual Absolute Differences in IQs

Comparison	Mean	Median	Range
$PrIQ_{IDS\text{-}2}$ vs. $ScrIQ_{IDS\text{-}2}$	7.94	7	0–37
$FSIQ_{IDS\text{-}2}$ vs. $ScrIQ_{IDS\text{-}2}$	7.00	6	0–39
$PrIQ_{IDS\text{-}2}$ vs. $FSIQ_{IDS\text{-}2}$	3.68	3	0–20
$FSIQ_{SB5}$ vs. $ScrIQ_{SB5}$	8.12	7	0–38
$FSIQ_{RIAS}$ vs. $ScrIQ_{RIAS}$	4.37	4	0–20

Note. IDS-2: N = 1,622; SB5: N = 1,829; RIAS: N = 2,109. PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ; IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation.

Table 3

Gamma Generalized Linear Models With Possible Predictors of Absolute Differences in IQs

	IDS-2	SB5	RIAS
Predictor	PrIQ vs. ScrIQ	FSIQ vs. ScrIQ	FSIQ vs. ScrIQ
Age	-0.00	-0.15 **	-0.08 *
Sex	0.00	0.03	0.03
AD(H)D	0.12	0.14	
Native language			
Bilingual	-0.14	0.00	-0.01
Other language	-0.04 **	0.07	-0.01
Education	0.06	0.05	0.05
IQ level			
Below-Average IQ	0.01 ***	0.07	0.27 ***
Above-Average IQ	0.06	0.26	0.17 **
Participation	0.00		
Cooperation		0.05	
Understanding		0.05	
Age*Below-Average IQ	-0.59 ***	0.12	-0.07
Age*Above-Average IQ	-0.10	0.31 **	-0.16
Likelihood	24.04 *	27.45 **	29.52 ***

Note. IDS-2: N = 1,566, SB5: N = 1,775, RIAS: N = 1,979. Displayed are regression coefficients standardized by dividing by two standard deviations (Gelman, 2008). Sex: 0 = male, 1 = female; AD(H)D: 0 = no, 1 = yes; Bilingual: 0 = German, 1 = bilingual; Other language: 0 = German, 1 = other native language; Education (of participants or their mothers): 0 = no university degree; 1 = university degree; Below average IQ: 0 = 85 $\leq$ IQ $\leq$ 115, 1 = IQ <85; Above average IQ: 0 = 85 $\leq$ IQ $\leq$ 115, 1 = IQ >115; Cooperation (in the test situation) and Understanding (of instructions): 0 = yes, 1 = partly/no. AD(H)D = Attention deficit/hyperactivity disorder or Attention deficit disorder; Participation = participation in the test situation; IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ.

*p <.05. **p <.01. ***p <.001.

Table 4

Reliabilities and Widths of 95% Confidence Intervals

			Internal consistency				Width of 95% CI				
IQ	Subtests[a]	Position[a]	Content overlap[a]	Overall[b]	Age-specific[b]	Age- and IQ-level-specific[c]	rtt[b]	95CI	$95CI_{age}$	$95CI_{ageIQ}$	$95CI_{rtt}$
$PrIQ_{IDS-2}$	14	1–14	1.00	.98	.95–.97	.67–.98	.85	8	10–13	8–28	21
$FSIQ_{IDS-2}$	7	1–7	.44	.97	.92–.95	.54–.97	.89	10	12–16	10–30	19
$ScrIQ_{IDS-2}$	2	6, 7	.38	.95	.83–.90	.53–.92	.86	13	17–23	15–30	21
$FSIQ_{SB5}$	10	1–10	.50	.99	.93–.98	.45–.96	.94	6	8–16	10–30	14
$ScrIQ_{SB5}$	2	1, 2		.97	.76–.93	.30–.93	.86	10	15–26	15–30	21
$FSIQ_{RIAS}$	4	1–4	1.00	.95	.93–.97	.55–.96	.88	13	10–16	11–30	19
$ScrIQ_{RIAS}$	2	1, 2		.93	.90–.94	.51–.96	.87	15	13–18	12–30	20

Note. IDS-2: N = 1,622; SB5: N = 1,829; RIAS: N = 2,109. Content overlap was calculated by dividing the number of subtests tapping the same broad abilities for both IQs by the total number of subtests over both IQs and concerns the respective IQ and the one in the row below (for the $ScrIQ_{IDS-2}$: the $PrIQ_{IDS-2}$). Mean test–retest intervals were 24 days (IDS-2), 22 days (SB5), and 19 days (RIAS). Age- and IQ-level-specific internal consistencies: IQ groups: <85, 85–115, >115; age groups: IDS-2: 5–6, 7–8, 9–12, 13–15, 16–20 years, SB5: <7, 7–8, 9–12, 13–15, 16–20, 21–29, 30–59, ≥ 60 years, RIAS: 3–4, 5–6, 7–8, 9–12, 13–15, 16–20, 21–59, ≥ 60 years. CI = confidence interval; rtt = test–retest reliability; 95CI = 95% CI with overall internal consistencies; $95CI_{age}$ = 95% CI with age-specific internal consistencies; $95CI_{ageIQ}$ = 95% CI with age- and IQ-level-specific internal consistencies; $95CI_{rtt}$ = 95% CI with test–retest reliability; PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ; IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation.

[a]Number of subtests per IQ and their position in the test sequence.

[b]Derived from manuals. Based on Cronbach's alphas (IDS-2 and RIAS) or split-half reliabilities (SB5).

[c]Computed according to the manuals with a formula provided by Lienert and Raatz (1998, p. 330) based on Cronbach's alphas (IDS-2 and RIAS) or split-half reliabilities (SB5).

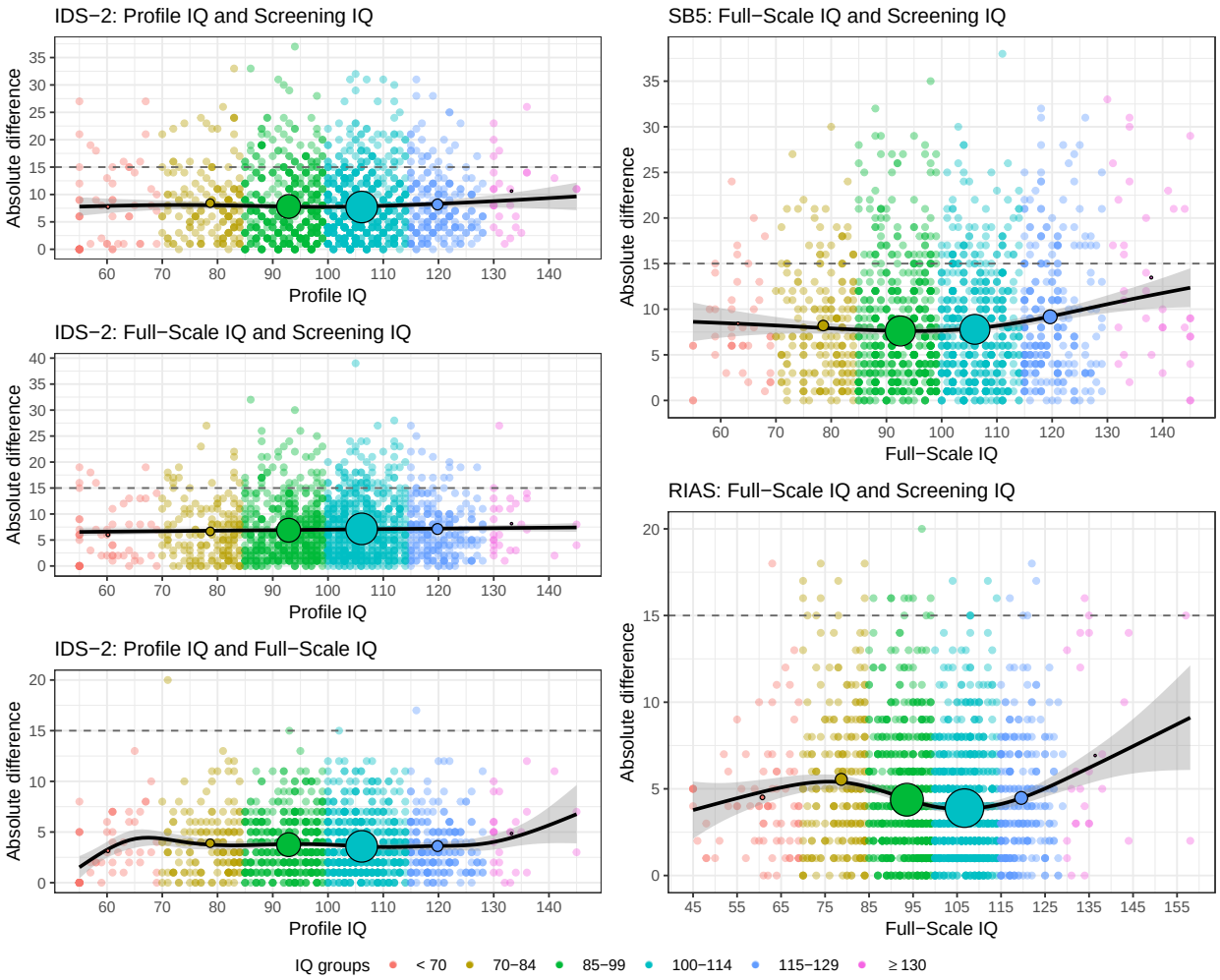

Figure 1. Intraindividual absolute differences between IQs within each test battery, plotted against the Profile IQ or Full-Scale IQ. The curve was fitted with general additive modeling and is displayed with a 95% confidence band. The dashed line indicates a difference of one standard deviation. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation.

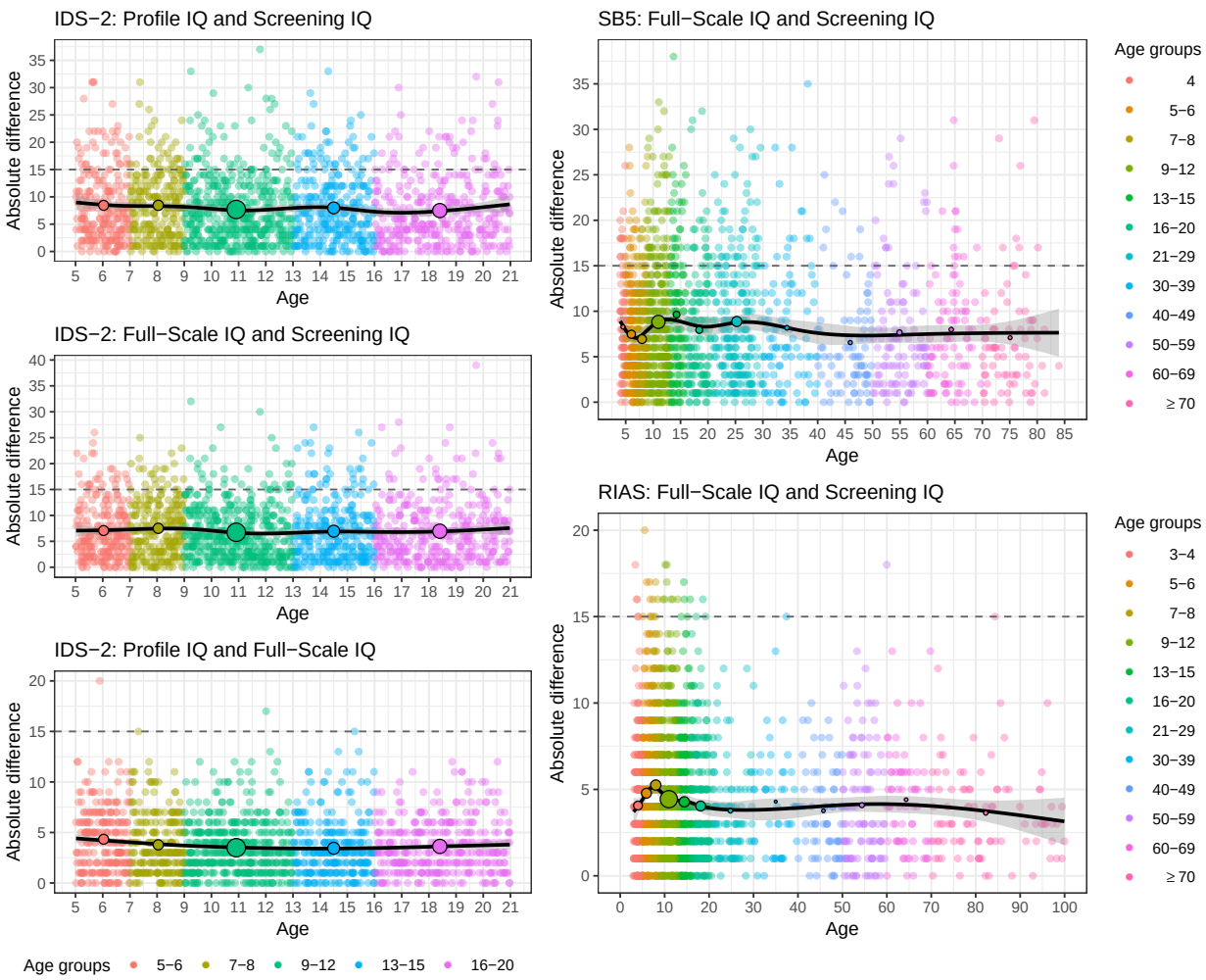

Figure 2. Intraindividual absolute differences between IQs within each test battery, plotted against age (in years). The curve was fitted with general additive modeling and is displayed with a 95% confidence band. The dashed line indicates a difference of 1 *SD*. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation.

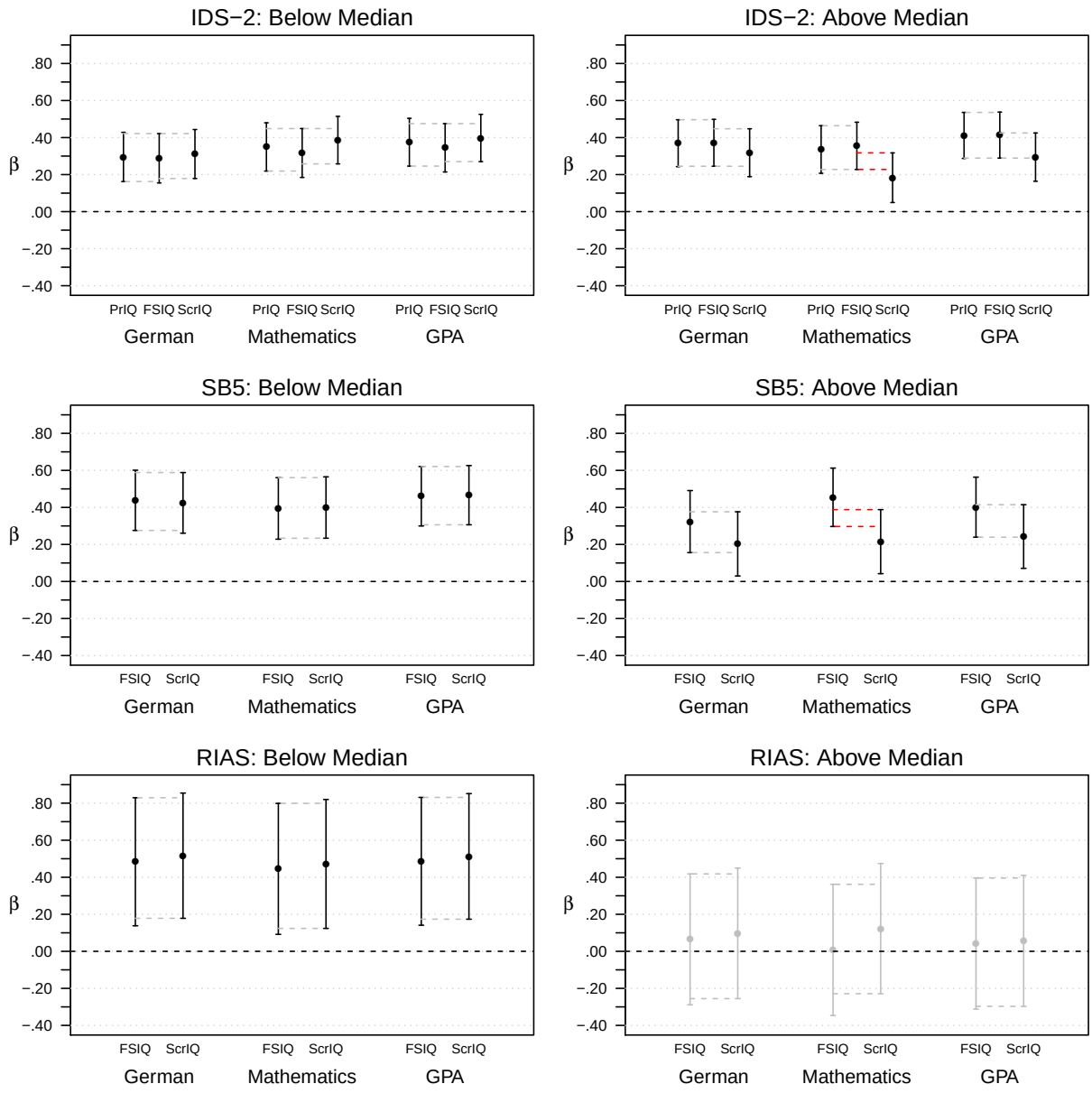

Figure 3. Comparison of 95% confidence intervals for standardized beta coefficients for IQs predicting school grades in German and mathematics and grade point average (GPA), splitted into subsamples with participants with intraindividual absolute differences in IQs below (IDS-2: $n = 203$, SB5: $n = 122$, RIAS: $n = 29$) and above (IDS-2: $n = 211$, SB5: $n = 127$, RIAS: $n = 35$) the median. A difference in betas was deemed significant if confidence intervals overlapped to a maximum of 50% (indicated in red). Significant betas are in black, nonsignificant betas in gray. Data for the IDS-2 and SB5 are cross-sectional; data for the RIAS are longitudinal. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ.

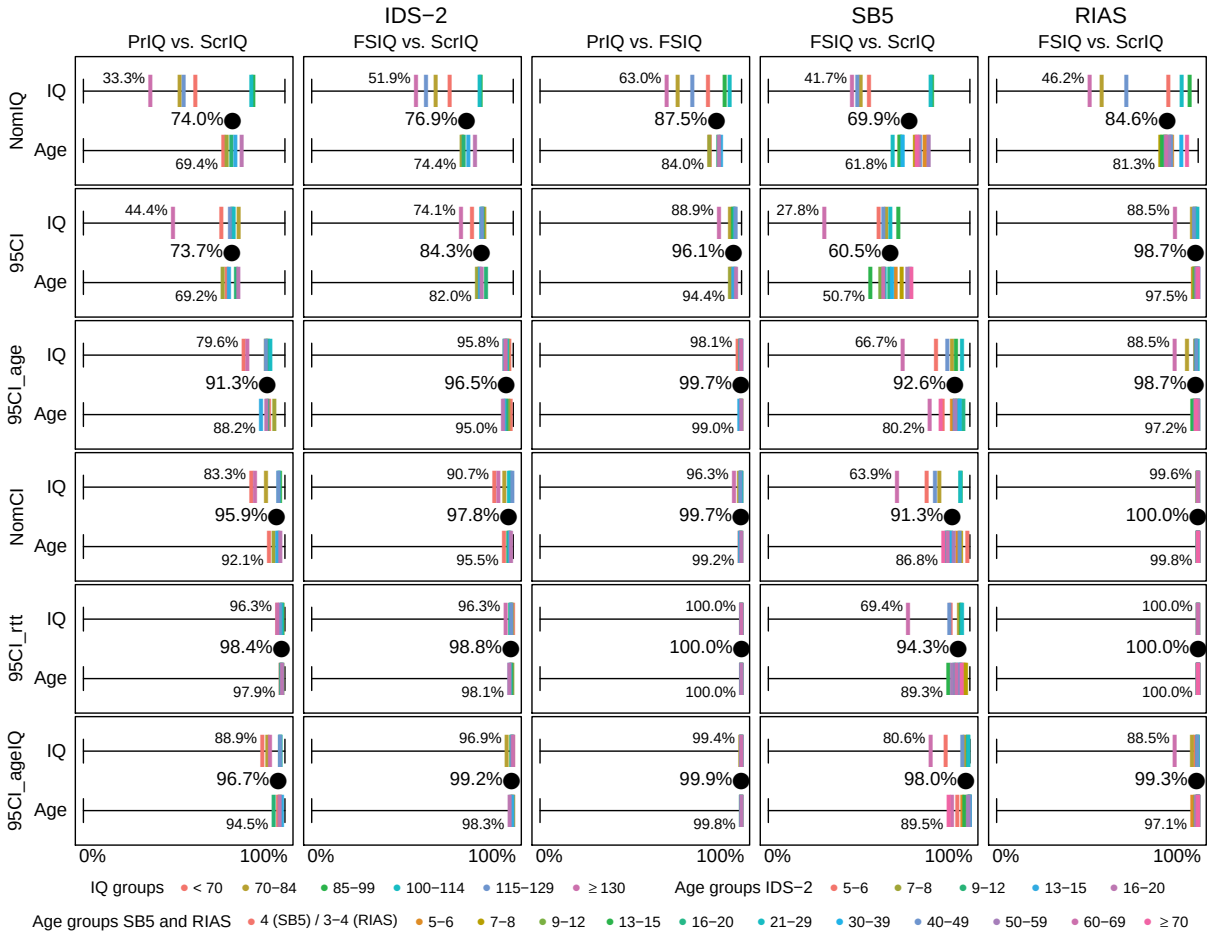

Figure 4. Percentage of participants with comparable IQs (i.e., overlapping intervals) determined by the following six criteria: NomIQ = IQ in nominal categories (e.g., "average" for IQ 85–115), 95CI = 95% CI with overall internal consistencies, $95CI_{age}$ = 95% CI with age-specific internal consistencies, NomCI = 95% CIs with overall internal consistencies in nominal categories, $95CI_{rtt}$ = 95% CI with test–retest reliabilities, and $95CI_{ageIQ}$ = 95% CI with age- and IQ-specific internal consistencies. The percentage of participants with comparable IQs both overall (black dots) and across IQ and age groups (color palette) are displayed. Numbers are displayed for the IQ and age group with the lowest percentage of participants with comparable IQs for each comparison. Ages given in years. IDS-2 = Intelligence and Development Scales–2; SB5 = Stanford–Binet Intelligence Scales–Fifth Edition, German adaptation; RIAS = Reynolds Intellectual Assessment Scales, German adaptation; PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ; CI = confidence interval.

APPENDIX C: Study III

Bünger, A., Urfer-Maurer, N., & Grob, A. (2019). Multimethod assessment of attention, executive functions, and motor skills in children with and without ADHD: Children's performance and parents' perceptions. *Journal of Attention Disorders*. Advance online publication. doi.org/10.1177/108705471882498

Article

Journal of Attention Disorders
1–11

DOI: 10.1177/1087054718824985
journals.sagepub.com/home/jad
SAGE

Multimethod Assessment of Attention, Executive Functions, and Motor Skills in Children With and Without ADHD: Children's Performance and Parents' Perceptions

Anette Bünger[1], Natalie Urfer-Maurer[1], and Alexander Grob[1]

Abstract

Objective: We examined whether children with attention-deficit/hyperactivity disorder (ADHD) differ from children without ADHD in attention, executive functions, and motor skills and whether measures of parents' perceptions and children's performance reveal comparable results. **Method:** About 52 children with ADHD and 52 children without ADHD aged 6 to 13 years completed performance-based measures of attention, executive functions, and motor skills. Parents completed questionnaires to rate their children's skills. **Results:** Parent questionnaires but not performance-based measures revealed higher inattention and lower executive function skills in children with ADHD compared to controls. For motor skills, both measurement methods revealed lower mean values and a higher number of children showing an impairment in the ADHD group. Parent-reported difficulties but not performance-based measures were related to the presence of an ADHD diagnosis. **Conclusion:** Our findings suggest that considering both parent questionnaires and performance-based measures will lead to a comprehensive picture of a child's strengths and difficulties. *(J. of Att. Dis. XXXX; XX[X] XX-XX)*

Keywords

ADHD, executive functions, motor skills, parent questionnaires, performance-based measures

Attention-deficit/hyperactivity disorder (ADHD) is one of the most often diagnosed neurobehavioral disorders in children with a worldwide prevalence of 5% in school-age children (Polanczyk, Silva de Lima, Horta, Biederman, & Rohde, 2007). Boys are diagnosed with ADHD two to four times more often than girls (Schmidt & Petermann, 2009). In addition to its core symptoms of inattention, hyperactivity, and impulsivity, ADHD is frequently accompanied by other symptoms, such as motor difficulties (Fliers et al., 2008), executive function deficits (Barkley, 1997; Hervey, Epstein, & Curry, 2004; Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005), and functional impairment, including learning problems (Loe & Feldman, 2007) and socioemotional difficulties (Maedgen & Carlson, 2000). These difficulties often accompany an ADHD diagnosis and are therefore here called accessory symptoms. The presence of early coexisting difficulties in motor development (Rasmussen & Gillberg, 2000) or executive functioning (Rinsky & Hinshaw, 2011) enhances the risk of poor psychosocial functioning. Hence examining potential accessory symptoms has been recommended (Subcommittee on Attention-Deficit/Hyperactivity Disorder Steering Committee on Quality Improvement and Management, 2011).

Although there are clear diagnostic criteria for ADHD in the international classification systems (*Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. [*DSM-5*]; American Psychiatric Association, 2013; *International Statistical Classification of Diseases and Related Health Problems*, 10th ed. [ICD-10]; World Health Organization, 1990), as well as guidelines to assess the aforementioned accessory symptoms, there is an ongoing debate on *how* to measure them, as different methods exist to assess the same symptoms. Clinical interviews, questionnaires, and performance-based measures (or their combination) are common methods to assess ADHD core and accessory symptoms. Therefore, it is necessary to know the advantages and disadvantages of each method and whether performance- and observer-based methods yield comparable results. This

[1]University of Basel, Switzerland

Corresponding Author:
Anette Bünger, Department of Psychology, Development and Personality Psychology, University of Basel, Missionsstrasse 62, 4055 Basel, Switzerland.
Email: anette.buenger@unibas.ch

question is the subject of the present study, which focused on attention as a core symptom and executive functions and motor skills as accessory symptoms of ADHD.

Questionnaires are time-saving, practical to use, and inexpensive. Moreover, they have the advantage of being able to assess the same symptoms in multiple respondents (Levy, Kronenberger, & Dunn, 2017). However, questionnaires reflect the respondent's perceptions and might therefore be subject to perspective biases (Pierrehumbert, Bader, Thévoz, Kinal, & Halfon, 2006), resulting in differences across informants who rate the same item. For instance, the correspondence between child-, parent-, and teacher reports was found to be low to modest with $r = .16$ to .52 for ADHD core and cognitive accessory symptoms (Caye, Machado, & Rohde, 2017; Duhig, Renk, Epstein, & Phares, 2000; Du Rietz et al., 2016; Pierrehumbert et al., 2006; Salwina et al., 2013) but strong with $r = .54$ for motor skills as accessory symptoms (Fliers et al., 2008). Among parents, disagreement can be found between mothers and fathers, with mothers tending to report more symptoms than fathers (Caye et al., 2017). Performance-based measures of attention, executive functions, and motor skills are more time consuming and more expensive compared to questionnaires. Nevertheless, a clear advantage of performance-based measures is the objectivity of results.

Both questionnaires and performance-based measures showed differences in attention, executive functions, and motor skills between children with and without ADHD (Fliers et al., 2008; Fried, Hirshfeld-Becker, Petty, Batchelder, & Biederman, 2015; Gioia, Isquith, Kenworthy, & Barton, 2002; Piek, Pitcher, & Hay, 1999): Regarding attention and executive functions, 15% to 89% of children with ADHD aged 7 to 13 years were found to be impaired (T score ≥ 65) on various attention and executive function skills based on the parent-rated Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000; Inhibit, 52%; Shift, 15%; Emotional Control, 26%; Initiate, 52%; Working Memory, 89%; Plan/Organize, 74%; Organization of Materials, 48%; Monitor, 63%; Gioia et al., 2002). Another study using seven subtests of the performance-based measure *Cambridge Neuropsychological Test Automated Battery (CANTAB;* Fried et al., 2015; *Verbal Recognition Memory, Intra-Extra Dimensional Set Shift [IED], Spatial Working Memory, Stockings of Cambridge [SOC], Reaction Time, Rapid Visual Information Processing [RVP], Affective Go/No-Go)* found 77% of children and adolescents with ADHD aged 8 to 15 years to have impairment (Z score < -1) on at least one attention and executive function measure, 52% on at least two measures, and 23% on at least three measures (Fried et al., 2015). Regarding motor skills, a study based on parent and teacher ratings on the Developmental Coordination Disorder Questionnaire (DCDQ; Kennedy-Behr, Wilson, Rodger, & Mickan, 2013) found one-third of children with ADHD aged 5 to 19 years to be impaired (raw score $<$ 10th percentile of normal controls) on the total DCD score (Fliers et al., 2008). Another study assessing motor skills with the performance-based Movement Assessment Battery for Children (MABC; Henderson & Sugden, 1992) found that 56% to 68% of children with ADHD aged 8 to 11 years showed problematic motor skills (total score $<$ 16th percentile) and 31% showed a severe general motor impairment (total score $<$ 5th percentile; Piek et al., 1999). Thus, it can be concluded that differential validity regarding ADHD is given for both methods. However, the abovementioned studies differ in definitions and cut-off values for impairments as well as in sample characteristics (e.g., age of subjects) and can therefore hardly be compared. Regarding attention and executive functions, previous studies showed that performance-based measures and parent ratings are not highly correlated although they both seem too reveal difficulties in ADHD children or other clinical samples (e.g., Anderson, Anderson, Northam, Jacobs, & Mikiewicz, 2002; Toplak, West, & Stanovich, 2013). However, less is known about the comparability of both methods for motor skills in ADHD children. Therefore, the current study addresses this issue by including attention, executive functions, and motor skills in one sample, applying comparable cut-off values for all three constructs.

As a first aim of the present study, we examined differences in parent-reported and performance-based attention, executive functions, and motor skills between children with and without ADHD (controls). In line with previous research, we hypothesized that children with ADHD would show more difficulties in attention, more problems in executive functions, and lower levels of motor skills compared to controls (group level; mean comparisons), and that these differences would emerge with both measurement methods (i.e., parent questionnaires and performance-based measures). Second, we hypothesized that the number of children showing an impairment in attention, executive functions, or motor skills would differ between children with and without ADHD (individual level; impairment comparisons). Assuming that parent questionnaires and performance-based measures are comparably valid measures to assess attention, executive functions, and motor skills, we finally explored the incremental predictive power of performance-based measures of attention, executive functions, and motor skills above parent questionnaire reports.

Method

Participants

The present study included 52 (out of 55) children with ADHD (M_{age} = 10.19 years, SD = 1.57; range: 7 to 13 years; 43 boys [83%], 9 girls [17%]) and 52 (out of 98) sex- and age-matched children without ADHD (control group; M_{age} = 10.13 years, SD = 1.64; range: 6 to 13 years; 43 boys [83%], 9 girls [17%]). All children attended

Table 1. Pearson Correlations for Attention and Executive Functions.

	Conners 3			CANTAB		
Scale	1	2	3	4	5	6
Conners 3						
1. Inattention		.73***	.02	.03	.06	.13
2. Executive functions	.53***/.57***		.08	−.06	.13	.03
CANTAB						
3. Sustained attention	.26/.03	.14/.26		−.13	.21	−.24*
4. Shifting and flexibility of attention	−.12/–.03	−.05/–.29	−.17/–.12		−.21	.30**
5. Spatial planning solved in minimum moves	.20/–.09	.25/–.01	.16/.26	−.14/–.29		−.22
6. Spatial working memory between errors	**–.25/.16**	−.13/–.10	−.27/–.23	.30/.22	−.18/−.30*	

Note. Values above the diagonal refer to the overall sample. Values before the solidus (/) refer to the attention-deficit/hyperactivity disorder group. Values after the solidus (/) refer to the control group. Significant differences between groups appear in bold. Data controlled for socioeconomic status, children's General Ability Index, age, and sex. Conners 3 = Conners Questionnaire for parents, 3rd edition; CANTAB = Cambridge Neuropsychological Test Automated Battery.
*$p < .05$. **$p < .01$. ***$p < .001$.

public primary school in Switzerland. Control children were recruited from local schools. Children with ADHD were recruited from and diagnosed by privately practicing pediatricians and the University Children's Hospital Basel. The diagnoses were based on *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) or *ICD-10* criteria and relied on parent interviews, questionnaires as well as psychological tests and medical checks to rule our other conditions. The presence or absence of an ADHD or any further diagnosis was confirmed by parents upon study enrollment. Three of 55 invited children with ADHD refused participation. Totally, 38 (73%) of the children with ADHD were on stimulant medication and discontinued ingestion 24 hr before testing, following the standard recommendation (Barkley, 1995; Thompson, 2007). None of the included children was diagnosed with autism spectrum disorder or met the criteria for intellectual disability (IQ < 70). (Co)-existing developmental disorders (e.g., learning disorders) are reported in Supplemental Table 1.

Procedure

Between May 2013 and May 2014 parents and children visited the laboratory at the University of Basel. Children completed performance-based measures of attention, executive functions, motor skills, and intelligence and parents were asked to complete questionnaires while the investigation of their children was taking place. Due to restrictions in testing time, a few children could not be assessed with all measures. The Ethics Committee of Northwest and Central Switzerland approved the study protocol. Written informed consent was obtained from parents for the children to participate and assent was obtained from each child.

Materials

Attention and executive functions

Parent questionnaire. Attention and executive functions were assessed using the short German version of the Conners Questionnaire for parents (3rd ed., Conners 3; Lidzba, Christiansen, & Drechsler, 2013). Parents were asked to rate the 42 items on a 4-point Likert-type scale ranging from 0 (*not at all/never*) to 3 (*very much/very frequently*). Higher scores indicate greater problems in attention and executive functions. Cronbach's alpha was .84 for the Executive Function scale and .90 for the Inattention scale.

Performance-based measure. Attention and executive functions were measured using four CANTAB (Fray, Robbins, & Sahakian, 1996) subtests. The Rapid Visual Processing (RVP; signal detection sensitivity) subtest assesses visual sustained attention, with higher scores indicating better sustained attention. The Intra-Extra Dimensional Set Shift (IED; mean errors) subtest assesses rule acquisition and reversal and shifting and flexibility of attention, with higher scores indicating lower flexibility of attention. The Stockings of Cambridge (SOC; problems solved in minimum moves) subtest measures spatial planning and problem solving, with higher scores indicating better planning skills. The Spatial Working Memory (between errors) subtest measures retention and manipulation of visuospatial information, with higher scores indicating lower working memory skills. Cronbach's alpha of the subtests are inconsistent, with high coefficients ($\alpha = .73$ to $.95$) reported by Luciana (2003) and low coefficients ($\alpha < .70$) reported by Syväoja and colleagues (2015).

Motor skills

Parent questionnaire. Parents were asked to rate their children's performance on motor activities using the German

version of the DCDQ (DCDQ-G; Kennedy-Behr et al., 2013). The DCDQ-G provides subscale scores in three dimensions: Control During Movement, Fine Motor and Handwriting, and General Coordination. Summed subscale scores yield a global DCD score. Parents were asked to rate 15 items using a 5-point Likert-type scale ranging from 1 (*not at all*) to 5 (*very much*). Higher scores indicate better motor skills. Cronbach's alpha of the scales ranged from .84 to .89 ($\alpha = .89$ for Control During Movement, $\alpha = .84$ for Fine Motor and Handwriting, $\alpha = .85$ for General Coordination, and $\alpha = .87$ for the global DCD score).

Performance-based measure. Motor skills were assessed using the German version of the MABC (2nd ed., MABC-2; Petermann, 2008). The MABC-2 contains eight subtests targeting three motor skills, manual dexterity, aiming and catching, and balance, and further provides a total motor score. Higher scores indicate better motor skills. According to Petermann (2008), Cronbach's alpha for the MABC-2 ranges from .41 to .62 ($\alpha = .52$ for manual dexterity, $\alpha = .41$ for aiming and catching, $\alpha = .46$ for balance, and $\alpha = .62$ for total motor score).

Control variables. Intelligence was measured using the General Ability Index (GAI) of the German version of the Wechsler Intelligence Scale for Children (4th ed., WISC-IV; Petermann & Petermann, 2011). The GAI consists of three verbal comprehension subtests and three perceptual reasoning subtests. Different from the Full-Scale IQ, the GAI does not include tasks targeting working memory processes and processing speed. The German Version of the WISC-IV meets standard psychometric properties (Daseking & Petermann, 2007). Cronbach's alpha for the WISC-IV GAI was $\alpha = .89$. To control for socioeconomic status (SES), parents were asked to indicate their highest level of education.

Statistical Analysis

In a preliminary step, we conducted partial correlations to test associations between parent questionnaires and performance-based measures of attention, executive functions, and motor skills, respectively. Then, to compare mean differences between children with ADHD and controls (Hypothesis 1), we conducted multivariate analyses of covariance for all measures of attention and executive functions (CANTAB; Conners 3) and motor skills (MABC-2; DCDQ-G). Pairwise comparisons were made using Bonferroni post hoc tests. These analyses were controlled for SES, children's intelligence, age, and sex.

Next, to examine impairment-based differences (comparing the number of children having an impairment) between children with ADHD and controls (Hypothesis 2), we conducted $\chi 2$ tests or Fisher's exact tests as indicated. We defined impairment on the Conners 3 scales, CANTAB tasks, and MABC-2 scales if a child scored below one standard deviation below the mean of the corresponding norm sample. For the DCDQ-G scales we defined impairment as any value below the 16th percentile of our control group (according to Fried et al., 2015), as they do not have norms.

Finally, to examine the incremental predictive power of performance-based measures compared to parent reports, we conducted multilevel binary logistic regressions for attention/executive functions as well as motor skills and with ADHD diagnosis as a dependent variable (ADHD diagnosis or no ADHD diagnosis). In Step 1, we entered the control variables SES and children's GAI, age, and sex. In Step 2, all parent-rated scales (independent variables) and in Step 3, all performance-based tasks (independent variables) were entered. Multilevel binary logistic regressions were conducted using age-standardized values for the Conners 3, CANTAB, and MABC-2, since they are generally used to set a diagnosis.

All tests were two-tailed, and statistical significance was defined at the .05 alpha level. We report eta squared (η^2) values for analyses of variance and odds ratios (ORs) for logistic regression analyses.

Little's missing completely at random test showed that missing values were not biased. Thus, we additionally imputed the control variables SES and GAI with the expectation maximization algorithm. The results did not change significantly when we imputed missing values. Therefore, and to enhance power, results with imputed data are reported below.

Results

Preliminary Analyses

Correlations for inattention and executive functions are reported in Table 1 and for motor skills in Table 2. Regarding inattention and executive functions, none of the parent-rated scales (Conners 3) were significantly related to the performance-based tasks (CANTAB), neither on the total sample level nor on the group level (ADHD and controls). Fisher's *z* transformation revealed that only 1 of 15 correlations (bold in Table 1) significantly differed between the ADHD and the control group. For motor skills, on a sample level most correlations between parent questionnaire scales (DCDQ-G) and performance-based tasks (MABC-2) were significant and ranged from weak to strong. Only the performance-based measure of manual dexterity was not significantly related to any of the parent-reported scales. On the group level, correlations were mostly stronger for the ADHD than the control group. However, Fisher's *z* transformation revealed that only 4 of 28 correlations (bold in Table 2) significantly differed between the two groups.

Table 2. Pearson Correlations for Motor Skills.

	DCDQ				MABC-2			
Scale	1	2	3	4	5	6	7	8
DCDQ								
1. Movement control		.53***	.81***	.90***	.15	.59***	.48***	.53***
2. Fine motor skills	.28/.55***		.63***	.80***	.14	.20	.18	.23*
3. General coordination	.81***/.70***	**.28/.70*** **		.93***	.21	.46***	.51***	.51***
4. Total score	91***/.84***	.57***/.86***	.89***/.92***		.19	.49***	.46***	.49***
MABC-2								
5. Manual dexterity	.12/–.20	−.12/.03	.05/.00	.04/–.05		.33**	.22*	.74***
6. Aiming and catching	**.67***/.19**	−.09/.04	.42**/.21	.48**/.17	.16/.42**		.36***	.76***
7. Balance	.42**/.37*	**–.20/.23**	.38*/.44**	.30/.40*	.03/.18	.36*/.19		.69***
8. Total score	**.58***/.06**	−.20/.10	.41**/.22	.39*/.15	.57***/.84***	.75***/.76***	.70***/.49***	

Note. Values above the diagonal refer to the overall sample. Values before the solidus (/) refer to the attention-deficit/hyperactivity disorder group. Values after the solidus (/) refer to the control group. Significant differences between groups appear in bold. Data controlled for socioeconomic status, children's General Ability Index, age, and sex. DCDQ =Developmental Coordination Disorder Questionnaire. MABC-2 = Movement Assessment Battery for Children, 2nd edition.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3. Mean-Level Differences Between the ADHD Group and Controls.

	ADHD ($n = 52$)	Controls ($n = 52$)			
Scale	*M* (*SD*)	*M* (*SD*)	*F*	*p* value	η2
Conners 3					
Inattention	10.13 (2.45)	4.36 (2.84)	88.25	<.001	.50
Executive functions	8.19 (3.30)	3.69 (2.84)	47.07	<.001	.35
CANTAB					
Sustained attention	0.94 (0.04)	0.96 (0.04)	0.839	.362	.01
Shifting and flexibility of attention	42.33 (21.41)	34.21 (20.92)	1.27	.263	.01
Spatial planning solved in minimum moves	7.17 (1.75)	7.45 (1.77)	0.03	.855	.00
Spatial working memory between errors	43.33 (17.71)	32.68 (17.65)	6.53	.012	.07
DCDQ					
Movement control	21.78 (5.19)	26.60 (3.02)	20.22	<.001	.19
Fine motor skills	11.84 (3.40)	17.22 (2.70)	51.46	<.001	.38
General coordination	15.04 (4.46)	21.38 (3.06)	47.07	<.001	.36
Total score	48.67 (10.81)	65.20 (7.53)	53.19	<.001	.39
MABC-2					
Manual dexterity	23.53 (5.93)	28.73 (5.38)	10.86	.001	.10
Aiming and catching	18.27 (4.93)	21.88 (4.50)	12.57	.001	.12
Balance	29.73 (5.47)	33.78 (2.63)	13.65	<.001	.13
Total score	71.55 (11.85)	84.39 (8.65)	25.68	<.001	.22

Note. Data controlled for socioeconomic status, children's General Ability Index, age, and sex. Conners 3 = Conners Questionnaire for parents, 3rd edition; CANTAB = Cambridge Neuropsychological Test Automated Battery; DCDQ = Developmental Coordination Disorder Questionnaire; MABC-2 = Movement Assessment Battery for Children, 2nd edition.

Mean Comparisons

Table 3 shows the mean differences between children with ADHD and controls (Hypothesis 1). Regarding attention and executive functions, parents reported higher inattention and lower executive function skills (Conners 3) for children with ADHD compared to children without ADHD with large effect sizes for both inattention ($\eta^2 = .50$) and executive functions ($\eta^2 = .35$). Regarding performance-based measures (CANTAB), children with ADHD made more between errors on spatial working memory tasks than the control group with results indicating a medium effect size ($\eta^2 = .07$). However, there were no significant group differences

Table 4. Comparisons of Number of Children Having an Impairment in the ADHD Group and the Control Group.

	ADHD	Controls			
Scale	*n* (%)	*n* (%)	χ^2	*p* value	η^2
Conners 3[a]					
Inattention	42 (87.5)	9 (20)	42.73	<.001	.46
Executive functions	30 (65.2)	8 (19.5)	18.41	<.001	.21
CANTAB[a]					
Sustained attention	12 (24.5)	7 (14.6)	1.51	.219	.02
Shifting and flexibility of attention	2 (4)	0 (0)	—	.495[b]	.02
Spatial planning solved in minimum moves	4 (8)	3 (6.3)	—	1.000[b]	.00
Spatial working memory between errors	4 (8.2)	3 (6)	—	1.000[b]	.00
DCDQ[c]					
Movement control	26 (56.5)	6 (13.3)	18.61	<.001	.20
Fine motor skills	31 (67.4)	6 (13.3)	27.56	<.001	.30
General coordination	31 (66)	6 (13.3)	26.48	<.001	.29
Total score	38 (80.9)	7 (15.6)	39.22	<.001	.43
MABC-2[a]					
Manual dexterity	25 (51)	9 7.(6)	12.40	<.001	.12
Aiming and catching	11 (22.4)	2 (3.9)	7.59	.006	.08
Balance	9 (18.4)	0 (0)	—	.001[b]	.10
Total score	18 (36.7)	2 (3.9)	16.82	<.001	.17

Note. Conners 3 = Conners Questionnaire for parents, 3rd edition; CANTAB = Cambridge Neuropsychological Test Automated Battery; DCDQ = Developmental Coordination Disorder Questionnaire; MABC-2 = Movement Assessment Battery for Children, 2nd edition.
[a]Impairment defined as 1 *SD* below the age-standardized average.
[b]Fisher's exact test.
[c]Impairment defined as value below the 16th percentile of the control group.

between children with ADHD and controls for all other measures of attention and executive functions, such as sustained attention, shifting and flexibility of attention, and spatial planning.

Regarding motor skills, parents reported significantly lower skills for children with ADHD on all subscales of the DCDQ-G with large effect sizes for movement control (η^2 = .19), fine motor skills (η^2 = .38), general coordination (η^2 = .36), and the total score (η^2 = .39). Moreover, children with ADHD had significantly lower scores on performance-based motor skill measures (MABC-2) than control children with medium effect sizes for manual dexterity (η^2 = .10), aiming and catching (η^2 = .12), and balance (η^2 = .13) and a large effect size for the total score (η^2 = .22).

Impairment Comparisons

Table 4 shows comparisons of the number of children having an impairment at $T \geq 60$ (1 *SD*) in the ADHD group and the control group (Hypothesis 2). Regarding parent-rated inattention (Conners 3), the number of children having an impairment was significantly higher in the ADHD group (87.5%) than in the control group (20%) with a large effect size of η^2 = .46. For parent-reported executive functions (Conners 3), the number of children having an impairment was significantly higher in the ADHD group (65.2%) than in the control group (19.5%) with a large effect size of η^2 = .21. However, when children had to perform tasks measuring attention and executive functions (CANTAB), the number of children having an impairment was comparable between the ADHD group (4% to 24.5%) and the control group (0% to 14.6%) on all tasks.

Regarding parent-reported motor skills (DCDQ-G), the number of children having an impairment was significantly higher in the ADHD group (56.5% to 80.9%) compared to the control group (13.3% to 15.6%) with all effect sizes in the large range (η^2 = .20 to .43). Also for performance-based measures of motor skills (MABC-2), the number of children having an impairment was significantly higher in the ADHD group (18.4% to 51%) than in the control group (0% to 6%) with medium effect sizes for manual dexterity (η^2 = .12), aiming and catching (η^2 = –.08), and balance (η^2 = .10) and a high effect size for the total score (η^2 = .17).

When applying stricter cut-off values for impairment definition, (e.g., 1.5 or 2 standard deviations), results for group comparison were almost identical for all attention and executive function measures (Conners 3 and CANTAB) as well as for parent-reported motor skills (DCDQ-G). Only for performance-based measures of motor skills (MABC-2), 4 out of 12 comparisons were not significant when considering also stricter cut-off values (percentile 9 and 5) what might be explained due to small case numbers (data not shown).

Table 5. Multilevel Binary Logistic Regression Comparing Parent Questionnaires and Performance-Based Measures of Attention and Executive Functions as Predictors of Attention-Deficit/Hyperactivity Disorder Diagnosis.

Regression step	*B*	*p* value	OR	95% CI for OR	
				OR_{min}	OR_{max}
Step 1					
SES	−0.22	.188	0.80	0.57	1.11
GAI	−0.06	.011	0.95	0.91	0.99
Sex	0.03	.959	1.03	0.31	3.45
Step 2					
Inattention[a]	0.31	.001	1.37	1.14	1.65
Executive functions[a]	0.10	.128	1.10	0.97	1.25
Step 3					
Sustained attention[b]	−1.06	.075	0.35	0.11	1.11
Shifting and flexibility of attention[b]	−1.17	.113	0.31	0.07	1.32
Spatial planning solved in minimum moves[b]	0.71	.426	2.02	0.36	11.46
Spatial working memory between errors[b]	−2.46	.050	0.085	0.01	1.00

Note. OR = Odds ratio; CI = confidence interval; OR_{min} = Lower value of the 95% CI for OR; OR_{max} = Higher value of the 95% CI for OR; SES = Socioeconomic status; GAI = General Ability Index; Conners 3 = Conners Questionnaire for parents, 3rd edition; CANTAB = Cambridge Neuropsychological Test Automated Battery.
[a]Conners 3 scales, age-standardized values.
[b]CANTAB tasks, age-standardized values.

Incremental Predictive Power of Performance-Based Measures Over Parent Reports

Tables 5 and 6 show the predictive power of attention and executive functions as well as motor skills for ADHD diagnosis. Regarding control variables, GAI was a significant negative predictor of an ADHD diagnosis with a small effect size of OR = .95. SES and children's age and sex did not significantly predict the presence of an ADHD diagnosis.

Regarding attention and executive functions, parent-reported inattention significantly predicted the presence of an ADHD diagnosis, with a small effect size of OR = 1.37. Parent-reported executive function did not significantly predict the presence of an ADHD diagnosis. Furthermore, none of the subsequently entered performance-based measures showed significant predictive power beyond that of parent questionnaires. Also, when performance-based measures of attention and executive functions were entered first in the statistical analyses (Step 1), none of these measures showed significant predictive power (data not shown).

Regarding motor skills, parent-reported fine motor skills and general coordination were significant negative predictors of an ADHD diagnosis, with small effect sizes for both fine motor skills (OR = .70) and general coordination (OR = .71). However, parent-reported movement control did not significantly predict the presence of an ADHD diagnosis. Furthermore, none of the subsequently entered performance-based measures showed significant predictive power beyond that of parent questionnaires.

Table 6. Multilevel Binary Logistic Regression Comparing Parent Questionnaires and Performance-Based Measures of Motor Skills as Predictors of Attention-Deficit/Hyperactivity Disorder Diagnosis.

Regression step	*B*	*p* value	OR	95% CI for OR	
				OR_{min}	OR_{max}
Step 1					
SES	−0.19	.262	0.83	0.60	1.15
GAI	−0.06	.011	0.95	0.91	0.99
Sex	0.11	.849	1.11	0.37	3.38
Step 2					
Movement control[a]	0.06	.666	1.06	0.82	1.37
Fine motor skills[a]	−0.36	.003	0.70	0.55	0.88
General coordination[a]	−0.31	.014	0.73	0.57	0.94
Step 3					
Manual dexterity[b]	−0.23	.171	0.80	0.58	1.10
Aiming and catching[b]	−0.22	.234	0.80	0.56	1.15
Balance[b]	−0.16	.372	0.85	0.60	1.22

Note. OR = Odds ratio; CI = confidence interval; OR_{min} Lower value of the 95% CI for OR; OR_{max} = Higher value of the 95% CI for OR; SES = Socioeconomic status; GAI = General Ability Index; DCDQ = Developmental Coordination Disorder Questionnaire; MABC-2 = Movement Assessment Battery for Children, 2nd edition.
[a]DCDQ scales, raw scores.
[b]MABC-2 tasks, age-standardized values.

When performance-based measures of motor skills were entered first in the statistical analyses (Step 1), only balance resulted as a significant negative predictor (B = –.24, p = .039) with a small effect size of OR = .79.

Discussion

The goal of this study was to shed light on differences in attention, executive functions, and motor difficulties between children with ADHD and controls and whether these symptoms are comparably revealed using different methods.

In line with Hypotheses 1 and 2, we found group differences regarding parent reports (Conners 3), with lower attention and executive functions in children with ADHD and more children showing an impairment in the ADHD compared to the control group. Our results are consistent with findings from Lidzba and colleagues (2013), who found differences in attention and executive functions based on the Conners 3 scales on a mean level as well as regarding the number of children having an impairment. In addition, we found significant mean-level differences on the CANTAB performance-based Spatial Working Memory subtest, which is in line with previous research that used other neuropsychological tasks and found working memory to be robustly impaired in children with ADHD (Martinussen, Hayden, Hogg-Johnson, & Tannock, 2005; Walshaw, Alloy, & Sabb, 2010; Willcutt et al., 2005). In contrast to our first hypothesis we did not find mean-level differences between the two groups regarding CANTAB performance-based sustained attention (RVP), shifting and flexibility of attention (IED), and spatial planning (SOC). Conversely, previous studies reported significant mean-level differences to the disadvantage of ADHD children for shifting and flexibility of attention (IED; Claesdotter, Cervin, Åkerlund, Råstam, & Lindvall, 2018; Fried et al., 2015; Gau & Shang, 2010) as well as for sustained attention (RVP) and spatial planning (SOC; Fried et al., 2015). Thus, they concluded that the CANTAB is a valuable instrument in the diagnostic assessment of ADHD. The inconsistency between our findings and previous results might be partially explained by differences in study design. Whereas Fried and colleagues (2015) and Claesdotter and colleagues (2018) tested children within a comparable age range to our study, they did not control for general cognitive ability. Gau and Shang (2010) did control for intelligence, but they examined older participants aged 11 to 17 years. Furthermore, inconsistent test reliability of CANTAB tasks could also partially explain divergent findings: Whereas Luciana (2003) reported high internal consistency coefficients (.73 to .95), Syväoja and colleagues (2015) found low internal consistency for most CANTAB tasks ($<$ .70), with both samples involving school-aged children.

Moreover, the number of children having an impairment did not differ between the ADHD and the control group regarding all four CANTAB measures, which is mostly in line with Fried and colleagues (2015). However, they found a higher number of children having an impairment on sustained attention in the ADHD group.

Our finding that differences in the number of impaired children are significant for parent reports but not for performance-based measures is also consistent with evidence from other traditional tests. McAuley, Chen, Goos, Schachar, and Crosbie (2010) concluded that a higher number of children show clinically relevant impairment using BRIEF parent reports compared to performance-based measures of executive functions. This finding seems to be independent of the type of clinical disorder (Drechsler & Steinhausen, 2013). A possible explanation for divergent findings between parent reports and performance-based measures may be that they do not fully measure the same skills of attention and executive functions. Whereas the Conners 3 questionnaire yields one general value of executive functions that also involves questions targeting behavioral and emotional aspects, the four performance-based CANTAB tasks differentiate between several subcomponents and exclusively target cognitive processes. In addition, questionnaires are more prone to reflect complex everyday situations whereas performance-based measures assess attention and executive functions in a more isolated manner due to the laboratory setting (Luciana, 2003). Specifically, diagnosticians try to set up optimal and unimpeded test situations, while parents might also think of challenging situations when completing questionnaires.

Regarding motor skills, we found significant group differences with worse motor skills and more children having an impairment in the ADHD compared to the control group for both the parent-rated DCDQ-G and the performance-based MABC-2 scales. These results are in line with prior research (Fliers et al., 2008; Piek et al., 1999).

Our results reveal that parent questionnaires and performance-based measures are more comparable for motor skills than for attention and executive functions, which was also revealed in stronger correlations between parent questionnaires and performance-based measures in motor skills than in attention and executive functions. In a similar vein, studies assessing the agreement between father and mother questionnaire ratings found higher agreement rates for behavior-based aspects such as hyperactivity than for more cognitive aspects such as inattention (Caye et al., 2017). Results derived from our study now add to this finding that there is also a higher agreement between questionnaires and performance-based measures for developmental behavioral symptoms compared to cognitive symptoms. This may be because motor skills occur in a more explicit behavior than cognitive abilities, which in turn are more internally oriented and manifest themselves in more subtle behavior (Caye et al., 2017). Therefore, it might be easier for parents to rate their children's motor skills more accurately than the less visible cognitive abilities. Another explanation might be that for motor skills, parent questionnaires and performance-based measures effectively target the same construct whereas this might not entirely be the case for attention and executive functions, as previously discussed.

Regarding our research question, we found no significant incremental predictive validity for performance-based measures of attention, executive functions, and motor skills compared to parent reports. Multilevel binary regression analysis showed that parent-reported difficulties in attention and motor skills (except movement control), but not performance-based measures, were related to the presence of an ADHD diagnosis. To understand this result, one has to bear in mind that *ICD-10* and *DSM-IV* criteria allow diagnosticians to base an ADHD diagnosis on questionnaires without using performance-based measures but not vice versa. Therefore, it is possible that diagnosticians more often relied on questionnaires including parent reports than on performance-based measures when giving the diagnosis, which could increase the predictive validity of parent questionnaires.

The inconsistent reliability of the CANTAB subtests could also have contributed to the fact that they did not show a predictive value for the presence of an ADHD diagnosis in our study. Yet, we used CANTAB as a performance-based measure for attention and executive functions because previous studies reported that the CANTAB subtests are suitable for determining deficits in attention and executive functions in children with ADHD (Fried et al., 2015; Gau & Shang, 2010). Moreover, there are still very few standardized test batteries in German-speaking countries that explicitly measure the executive functions of school-aged children, even though research shows how important the concept of executive functions is for academic achievement as well as health (Best, Miller, & Naglieri, 2011; Moffitt et al., 2011), especially in the context of an ADHD diagnosis (Willcutt et al., 2005). In 2018, the Intelligence and Development Scales–2 (Grob & Hagmann-von Arx, 2018), a newly developed performance-based paper-and-pencil test battery measuring attention and executive functions, became available for German-speaking children and adolescents. Future research examining other parent questionnaires and performance-based measures will reveal whether the lack of correspondence is a general issue or caused by a possible bias of parent reports and/or possible unreliability of computerized tests. Finally, we suggest that it may be essential for practitioners to know the children's performance in a standardized test situation beyond the symptoms reported by the child, the parents, or others in order to gain a comprehensive picture of the strengths and difficulties of a child, to design an individualized treatment plan, and to draw direct comparisons between children.

Limitations

Our study has some limitations. First, we relied on the diagnosis given by experienced practicing pediatricians and were therefore not able to control for the standardization of the diagnostic process. Second, the interpretation of questionnaires is limited to parental perceptions of children's attention, executive functions, and motor skills, as teacher ratings and the child's perspective were not included in this study. Third, we focused on attention, executive functions, and motor skills, but there are other coexisting symptoms, such as socioemotional difficulties, that often accompany an ADHD diagnosis and play an important role in psychological well-being. Finally, for performance-based measures of attention and executive functions we used only computer-based tasks. Future research could also include multimethod tests of socioemotional difficulties as well as performance-based measures of executive functions and attention that are based on paper-and-pencil or interaction tasks.

Conclusion

In sum, children with ADHD have more difficulties in attention, executive functions, and motor skills and are more often impaired in these skills compared to controls. Whereas parent questionnaires and performance-based measures reveal comparable results regarding differences in motor skills between children with ADHD and controls, only parent questionnaires reveal differences in attention and executive functions between the two groups. We propose applying parent questionnaires as well as performance-based measures in order to better understand a child's needs and to design an individualized treatment plan, since it is possible that parent questionnaires and performance-based measures regarding attention and executive functions do not completely target the same construct.

Acknowledgments

The authors sincerely thank Dr. Priska Hagmann-von Arx and Dr. Sakari Lemola for their valuable comments on this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Anette Bünger received financial support from the Suzanne und Hans Biäsch Stiftung zur Förderung der angewandten Psychologie during the preparation of the paper.

Supplemental Material

Supplemental material for this article is available online.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Anderson, V. A., Anderson, P., Northam, E., Jacobs, R., & Mikiewicz, O. (2002). Relationships between cognitive and behavioral measures of executive function in children with brain disease. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *8*, 231-240. doi:10.1076/chin.8.4.231.13509

Barkley, R. A. (1995). *Taking charge of ADHD: The complete, authoritative guide for parents*. New York, NY: Guilford Press.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*, 65-94. doi:10.1037/0033-2909.121.1.65

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*, 327-336. doi:10.1016/j.lindif.2011.01.007

Caye, A., Machado, J. D., & Rohde, L. A. (2017). Evaluating parental disagreement in ADHD diagnosis: Can we rely on a single report from home? *Journal of Attention Disorders*, *2*, 561-566. doi:10.1177/1087054713504134

Claesdotter, E., Cervin, M., Åkerlund, S., Råstam, M., & Lindvall, M. (2018). The effects of ADHD on cognitive performance. *Nordic Journal of Psychiatry*, *72*, 158-163. doi:10.1080/08039488.2017.1402951

Daseking, M., & Petermann, U. (2007). Neue diagnostische Verfahren Intelligenzdiagnostik mit dem HAWIK-IV [Assessment of intelligence with the HAWIK-IV]. *Kindheit Und Entwicklung*, *16*, 250-259. doi:10.1026/0942-5403.16.4.250

Drechsler, R., & Steinhausen, H. C. (2013). *Verhaltensinventar zur Beurteilung exekutiver Funktionen BRIEF. Deutschsprachige Adaption des Behavior Rating Inventory of Executive Function* [BRIEF. Behavior Rating Inventory of Executive Function. German Adaptation]. Bern, Switzerland: Huber.

Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice*, *7*, 435-453. doi:10.1093/clipsy.7.4.435

Du Rietz, E., Cheung, C. H. M., McLoughlin, G., Brandeis, D., Banaschewski, T., Asherson, P., & Kuntsi, J. (2016). Self-report of ADHD shows limited agreement with objective markers of persistence and remittance. *Journal of Psychiatric Research*, *82*, 91-99. doi:10.1016/j.jpsychires.2016.07.020

Fliers, E., Rommelse, N., Vermeulen, S. H., Altink, M., Buschgens, C. J., Faraone, S. V., . . . Buitelaar, J. K. (2008). Motor coordination problems in children and adolescents with ADHD rated by parents and teachers: Effects of age and gender. *Journal of Neural Transmission*, *115*, 211-220. doi:10.1007/s00702-007-0827-0

Fray, P., Robbins, T. W., & Sahakian, B. J. (1996). Neuropsychiatric applications of CANTAB. *International Journal of Geriatric Psychiatry*, *11*, 329-336. doi:10.1002/(SICI)1099-1166(199604)11:4%3C329::AID-GPS453%3E3.0.CO;2-6

Fried, R., Hirshfeld-Becker, D., Petty, C., Batchelder, H., & Biederman, J. (2015). How informative is the CANTAB to assess executive functioning in children with ADHD? A controlled study. *Journal of Attention Disorders*, *20*, 1-8. doi:10.1177/1087054712457038

Gau, S. S., & Shang, C. (2010). Executive functions as endophenotypes in ADHD: Evidence from the Cambridge Neuropsychological Test Battery (CANTAB). *The Journal of Child Psychology and Psychiatry*, *7*, 838-849. doi:10.1111/j.1469-7610.2010.02215.x

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Test review: Behavior Rating Inventory of Executive Function. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *6*, 235-238. doi:10.1076/chin.6.3.235.3152

Gioia, G. A., Isquith, P. K., Kenworthy, L., & Barton, R. M. (2002). Profiles of everyday executive function in acquired and developmental disorders. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *8*, 121-137. doi:10.1076/chin.8.2.121.8727

Grob, A., & Hagmann-von Arx, P. (2018). *Intelligence and Development Scales–2. Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche (IDS-2)* [IDS-2. Intelligence and Development Scales for Children and Adolescents]. Bern, Switzerland: Hogrefe.

Henderson, S. E., & Sugden, D. A. (1992). *Movement assessment battery for children*. Kent, UK: The Psychological Corporation.

Hervey, A. S., Epstein, J. N., & Curry, J. F. (2004). Neuropsychology of adults with attention-deficit/hyperactivity disorder: A meta-analytic review. *Neuropsychology*, *18*, 485-503. doi:10.1037/0894-4105.18.3.485

Kennedy-Behr, A., Wilson, B. N., Rodger, S., & Mickan, S. (2013). Cross-cultural adaptation of the Developmental Coordination Disorder Questionnaire 2007 for German-speaking countries: DCDQ-G. *Neuropediatrics*, *44*, 245-251. doi:10.1055/s-0033-1347936

Levy, J. D., Kronenberger, W. G., & Dunn, D. W. (2017). Development of a very brief measure of ADHD: The CHAOS scale. *Journal of Attention Disorders*, *2*, 575-586. doi:10.1177/1087054713497792

Lidzba, K., Christiansen, H., & Drechsler, R. (2013). *Conners Skalen zu Aufmerksamkeit und Verhalten-3. Deutschsprachige Adaptation der Conners 3rd edition von Keith Conners* [German adaptation and normalization of the Conners-3 questionnaires]. Bern, Switzerland: Huber.

Loe, I. M., & Feldman, H. M. (2007). Academic and educational outcomes of children with ADHD. *Ambulatory Pediatrics: The Official Journal of the Ambulatory Pediatric Association*, *7*, 82-90. doi:10.1016/j.ambp.2006.05.005

Luciana, M. (2003). Practitioner review: Computerized assessment of neuropsychological function in children: Clinical and research applications of the Cambridge Neuropsychological Testing Automated Battery (CANTAB). *The Journal of Child Psychology and Psychiatry*, *44*, 649-663. doi:10.1111/1469-7610.00152

Maedgen, J. W., & Carlson, C. L. (2000). Social functioning and emotional regulation in the attention deficit hyperactivity disorder subtypes. *Journal of Clinical Child Psychology*, *29*, 30-42. doi:10.1207/S15374424jccp2901

Martinussen, R., Hayden, J., Hogg-Johnson, S., & Tannock, R. (2005). A meta-analysis of working memory impairments in

children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 377-384. doi:10.1097/01.chi.0000153228.72591.73

McAuley, T., Chen, S., Goos, L., Schachar, R., & Crosbie, J. (2010). Is the Behavior Rating Inventory of Executive Function more strongly associated with measures of impairment or executive function? *Journal of the International Neuropsychological Society*, *16*, 495-505. doi:10.1017/S1355617710000093

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 2693-2698. doi:10.1073/pnas.1010076108

Petermann, F. (2008). *Movement Assessment Battery for Children* (2nd ed.). Frankfurt, Germany: Pearson Assessment.

Petermann, F., & Petermann, U. (Eds.). (2011). *Wechsler Intelligence Scale for Children (WISC-IV)*. Frankfurt, Germany: Pearson Assessment.

Piek, J. P., Pitcher, T. M., & Hay, D. A. (1999). Motor coordination and kinaesthesis in boys with attention deficit-hyperactivity disorder. *Developmental Medicine & Child Neurology*, *41*, 159-165. doi:10.1111/j.1469-8749.1999.tb00575.x

Pierrehumbert, B., Bader, M., Thévoz, S., Kinal, A., & Halfon, O. (2006). Hyperactivity and attention problems in a Swiss sample of school-aged children: Effects of school achievement, child gender, and informants. *Journal of Attention Disorders*, *10*, 65-76. doi:10.1177/1087054705286050

Polanczyk, G. B., Silva de Lima, M., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: A systematic review and metaregression analysis. *American Journal of Psychiatry*, *164*, 942-948. doi:10.1176/ajp.2007.164.6.942

Rasmussen, P., & Gillberg, C. (2000). Natural outcome of ADHD with developmental coordination disorder at age 22 years: A controlled, longitudinal, community-based study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*, 1424-1431. doi:10.1097/00004583-200011000-00017

Rinsky, J. R., & Hinshaw, S. P. (2011). Linkages between childhood executive functioning and adolescent social functioning and psychopathology in girls with ADHD. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *17*, 368-390. doi:10.1080/09297049.2010.544649

Salwina, W., Ismail, W., Baharudin, A., Ruzyanei, N., Jaafar, N., & Midin, M. (2013). Attention deficit hyperactivity disorder symptoms reporting in Malaysian adolescents: Do adolescents, parents and teachers agree with each other? *Asian Journal of Psychiatry*, *6*, 483-487. doi:10.1016/j.ajp.2013.05.001

Schmidt, S., & Petermann, F. (2009). Developmental psychopathology: Attention deficit hyperactivity disorder (ADHD). *BMC Psychiatry*, *9*, Article 58. doi:10.1186/1471-244X-9-58

Subcommittee on Attention-Deficit/Hyperactivity Disorder Steering Committee on Quality Improvement and Management. (2011). ADHD: Clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics*, *128*, 1007-1022. doi:10.1542/peds.2011-2654

Syväoja, H. J., Tammelin, T. H., Ahonen, T., Räsänen, P., Tolvanen, A., Kankaanpää, A., & Kantomaa, M. T. (2015). Internal consistency and stability of the CANTAB neuropsychological test battery in children. *Psychological Assessment*, *27*, 698-709. doi:10.1037/a0038485

Thompson, K. (2007). *Medicines for mental health: The ultimate guide to psychiatric medication*. North Charleston, SC: Book Surge.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *The Journal of Child Psychology and Psychiatry*, *54*, 131-143. doi:10.1111/jcpp.12001

Walshaw, P. D., Alloy, L. B., & Sabb, F. W. (2010). Executive function in pediatric bipolar disorder and attention-deficit hyperactivity disorder: In search of distinct phenotypic profiles. *Neuropsychology Review*, *20*, 103-120. doi:10.1007/s11065-009-9126-x

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry*, *57*, 1336-1346. doi:10.1016/j.biopsych.2005.02.006

World Health Organization. (1990). *The international statistical classification of diseases and related health problems* (10th ed.). Geneva, Switzerland: Author.

Author Biographies

Anette Bünger is a Consultant at the School Psychology Service in Basel, and PhD candidate in the graduate program school psychology, developmental diagnostics and educational counseling (SEED) integrated in the Department of Development and Personality Psychology at University of Basel.

Natalie Urfer-Maurer, PhD, is a Postdoctoral Researcher in the Department of Development and Personality Psychology at University of Basel.

Alexander Grob, PhD, is a Professor of Development and Personality Psychology (chair) at the University of Basel and head of the graduate degree program in school psychology, developmental diagnostics and educational counseling (SEED).